DOCUMENT RESUME

ED 098 814 FL 006 605

AUTHOR	DeRocher, James E.; And Others
TITLE	The Counting of Words: A Review of the History, Techniques and Theory of Word Counts with Annotated Bibliography.
INSTITUTION	Syracuse Univ. Research Corp., N.Y.
SPONS AGENCY	Defense Language Inst., Washington, D.C.
REPORT NO	AD-775-922; SURC-TR-73-177
PUB DATE	May 73
NOTE	302p.
AVAILABLE FROM	National Technical Information Service, Springfield, Virginia 22151 (Order No. AD-775 922, MF-$1.45; HC-$18.00)
EDRS PRICE	MF-$0.75 HC-$15.00 PLUS POSTAGE
DESCRIPTORS	Annotated Bibliographies; Computational Linguistics; Contrastive Linguistics; Descriptive Linguistics; Diachronic Linguistics; Etymology; *Language Research; *Mathematical Linguistics; *Structural Analysis; Vocabulary; *Word Frequency; *Word Lists

ABSTRACT

As part of a continuing project of language analysis, SURC presents an essay on the nature and history of frequency counts. The first section deals with the history of such counts and traces them from Early Hellenic times to the present. Section 2 is an analysis of techniques used and describes the capabilities and limitations of frequency counts taken in both English and foreign languages. Section 3 is an analysis of the statistical lawfulness of vocabulary distributions and presents a comparison and evaluation of the theoretical models used to describe vocabulary distributions. Section 4 is an annotated bibliography with an author index provided. (Author)

BEST COPY AVAILABLE

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

FL 006605

The Counting of Words:

A Review of the History, Techniques and Theory of Word Counts With Annotated Bibliography

Prepared for
The Defense Language Institute
Under Contract DAAG05-72-C-0578

by
James E. DeRocher
Murray S. Miron
Sam M. Patten
Charles C. Pratt

May 1973

2

SURC

TABLE OF CONTENTS

Page No.

SECTION I

Narrative. 1

SECTION II

Discussion 58

SECTION III

Analyses of the Statistical Lawfulness of Vocabulary Distributions. 104

SECTION IV

Bibliography 185

THE HISTORY, TYPES, AND PROCEDURES OF FREQUENCY COUNTS

SECTION I - NARRATIVE

On a world scale, the counting of the frequency of occurrence of linguistic elements has had a long history motivated by diverse purposes. Traditionally the counts have been of words, but they have also been of phonemes, morphemes, syllables or idiomatic expressions. The purpose of these counts has usually been to develop a vocabulary of a special type such as of rare, frequent, useful, or important words with the ultimate objective of developing vocabularies for the teaching and learning of stenography, spelling, or reading in the easiest and most efficient manner possible.

The counts with which we are familiar extend far back into the Mediterranean world, where, in one instance, we find that the scholars of Alexandria (Egypt) distinguished between rare and frequent words of Homeric Greek for the benefit of local students of Literary Greek. In the Tenth Century, the Talmudists categorized and counted the words in the Torah.

In general, the West European history of frequency counts proceeds from a similar implicit assumption that the best way to learn a language, either your own, or someone else's is to know which words you will meet most often so that you can learn them first. Earlier lists tended to be limited beginner's vocabularies usually of "useful, hard words". Such lists were compiled as early as the Fifteenth Century.

Other early lists tended to be for occupational or instructional purposes providing mainly name lists such as of birds, animals, parts of the body and of occupations.

These early lists tended to be restricted, special purpose vocabularies, and it wasn't until 1721 that Nathanial Bailey attempted to compile an extensive English vocabulary to serve as the basis for a dictionary.

In 1588, Timothe Bright published his "Characterie: An Arte of Shorte, Swifte, and Secrete Writing by Character". It was the first known attempt at developing a form of shorthand, although it was not phonetically based. It was also the first attempt at developing a self contained, basic or "Island" vocabulary capable of being used to express all necessary concepts with as few words (or symbols) and their variations as possible.

In 1649, Sulanus published a list of words appearing in Homeric Greek and a little later a cleric named Winckler who lived in Hamburg annotated a Greek-German version of the New Testament to indicate words occurring only once or only in a single verse. A similar but more elaborate Dutch-Greek version of that annotation published in 1698 over the signature of Johannes Leusden indicated that there were 4,956 different words in the Dutch version of the New Testament of which 1,686 occurred only once or only in one verse. He marked all those that occurred only once, all those that occurred twice or more, and indicated the number of new words occuring in each chapter.

During the Sixteenth and Seventeenth Centuries various counts of the vocabulary in literature were published in England, including one of the Authorized Version of the New Testament which in English was determined to contain 6,000 different words. However, we do not know the principles on which the counts were made; i.e., whether only headwords were counted, or whether derivatives and inflections were also counted. The same uncertainty applies to the counts of this era of Shakespeare's plays (21,000 words) and Milton's (8,000 different words in Paradise Lost).

Nineteenth Century Counts

In the last half of the Nineteenth Century, two frequency counts of different types appeared which anticipated the proliferation of counts which has taken place in this century. The first was W. D. Whitney's phonetic study on the relative frequency of 10,000 sounds found in 10 classics of English literature as drawn from 1,000 sound samples from each of five prose and five poetic works. The word count type of study of this era is best exemplified by F. W. Kaeding's count.

Kaeding's work was the first and one of the largest of the modern frequency counts. With the assistance of nearly 6,000 contributors and co-workers, Kaeding collected and counted some 11 million German words and 20 million syllables from 14 categories of material. He found 258,173 different words of which half occurred only once. The purpose of the count was to assist in teaching stenography rather than language, so homonyms were listed only once regardless of word meaning although derivatives and inflections were listed separately. In German, this meant some physical separation in the alphabetical summary lists because of the phonetic spelling of umlauted letters; plurals being separated from their singulars and verb forms scattered, making the determination of the total frequency of a semantic concept a task involving much hunting for related word forms and adding of frequencies. (Morgan subsequently corrected many of these shortcomings in his revision published in 1928.) Nevertheless, this work is important not only for its size and detail, but because it firmly established the method of counting large numbers of words from a wide variety of sources in order to find truly general or representative words, and established frequency of occurrence of a word as the basis for a determination of its linguistic importance or value. This use of frequency is more valid for shorthand word lists than for other purposes, and has since been modified by later researchers preparing vocabularies for other purposes.

Before the close of the Nineteenth Century, there appeared a more modest word count in English. It was J. M. Rice's "Rational Spelling

Book". It is important mainly because it reveals another reason for frequency counting; the teaching of spelling of the real word, as opposed to some representation of it as in shorthand. The frequency list was designed to determine which words were used most often and, therefore, those which should be learned first. In spelling as well as stenography in which the word is more important than its meaning, frequency can be expected to adequately index the importance of a word.

1904-1920

In 1904, Reverend J. Knowles, in England, while developing his London Point System of Reading for the Blind, made a frequency count of literature, principally from the Bible, for a total of 100,000 running words. From this corpus he derived a list of 350 most frequently occurring words and indicated the frequency of each. This count is of interest because it was among the first to note that the first noun appeared as number 73 in order of frequency rank. The preceding 72 were the so-called structural, semi-structural, grammatical or relating words. This phenomenon of the late appearance of nouns, resulting from the use of raw frequency data to determine the importance of a word for purposes of vocabulary compilation is, in fact, inherent in frequency counts and has led later word counters to use various procedures to compensate for the disproportionate dominance of the grammatical functors. Keniston and Thorndike were among the first to make such modifications.

In 1910, W. E. Chancellor wrote an article entitled "Spelling: 1,000 Words". He had derived his 1,000 words from some 20,000 that he compiled and then screened. His work is important not because of the techniques which he used, which were neither particularly well documented nor sophisticated but because he employed personal correspondence rather than literature as his source material. Later investigators using better techniques have made excellent use of such sources.

In 1911, R. C. Eldridge made a study of newspaper language in the Buffalo, New York area and published the results of his study as "Six Thousand Common English Words." He made the study for the purpose of developing a limited universal vocabulary. His courses didn't lead to much universality, but his work is cited because it: 1) involved counting only 44,000 running words (which really are not enough to be statistically sound, much less universal) and 2) involved only one broad category of material, i.e., newspaper English from four newspapers in one localized area. However, the count has been used as late as 1967 for purposes of comparison by Beier, Starkweather, and Miller in their "Analysis of Word Frequencies in Spoken Language of Children" (1967), a study they made of the oral vocabulary of grade school children in Salt Lake City.

In the 1913-15 period, Leonard Ayres compiled and published two works: "The Spelling Vocabularies of Personal and Business Letters" and "Measuring Scale for Ability in Spelling". He assessed some 400,000 running words from 2500 people in 75 communities. About 70 percent of the material was from personal and business letters. The data collected from these sources were used to devise his "Measuring Scale for Ability in Spelling". This may have been the first use of word counts for testing of students. Ayres' work clearly demonstrated the fact that while languages have many words, the frequencies of usage of these words are heavily skewed; an absolutely small number of word types accounting for the majority of the words used in ordinary discourse. For example, he found that his 50 commonest words made up 50 percent of the materials he counted; 300 words (in order of frequency) made up 75 percent, and 1,000 made up 90 percent. Later studies in confirming this stable finding stimulated interest in basic and limited (but universal) vocabularies.

Cook and O'Shea in 1914 made their "Concrete Investigation of the Material of English Spelling" based on the family letters of 13 informants. Cook and O'Shea found 5,200 different words in the 200,000 they counted. Allowing for the small corpus and limited range of the samples, the results as far as word usage is concerned replicated Ayres' findings on the heavy use of a few words, in particular the function words. They found that the first nine words on their frequency list constituted 25 percent of all words in the corpus, and the first 42 more than 50 percent, all of which were functors. Of the remainder, 963 words included 91 percent of all running words even though another 4237 were used one or more times for the remaining 9 percent of word usage.

In 1920, Hayward Keniston published one of the earliest of the foreign language counts prepared for the purpose of second language instruction. It was called "Common Words in Spanish" and is particularly noteworthy because Keniston was apparently the first vocabulary compiler to recognize that word value or importance does not depend on word occurrence frequency alone. It must, therefore, depend on some other factor or factors related to the uses to which the list is to be put. Keniston apparently believed that if representative or general-use words are desired, the number of different sources in which a word appears might be as important as how often the word appears. If the occurrence of a word was restricted to relatively few sources or types of sources, Keniston argued that such restriction indicated that the word was either peculiar to the author or to the subject matter and, hence, had lower value even though its frequency in such sources might be high. In recognition of this fact, Keniston noted the effect of range (or number of sources in which a word appeared) by using two lists. The lists were based on frequency of occurrence, but in one he placed only those words which qualified by appearing in 80 percent of his sources, and in the other he listed only those which appeared in at least 66 percent of his sources. These sources, incidentally, were mainly plays, but included newspapers, reviews, short stories, and novels (all printed sources, except insofar as pseudo-oral material might be included, particularly in plays, speeches or in quotations).

The Early Modern Counts--1920-1930

In 1921, Edward Thorndike's seminal counts of English first appeared and heralded an era of his authority which was to last until today. Like Kaeding's count of German, Thorndike for the first time had carefully described the word frequencies of a massive corpus of the written language.

In spite of criticism by Dewey, Palmer, Bongers, Rosenzweig and McNeil among others, his works have survived. Thorndike, an educational psychologist, was probably the greatest exponent of objective frequency counts as the basis for teaching, principally in the areas of reading, vocabulary building, and spelling, although his Teacher's Word Lists have also been used for preparing graded textbooks and achievement tests. His first "Teachers' Wordbook of 10,000 Words" which appeared in 1921 was successively revised in 1931 to 20,000 words and in 1944 (in collaboration with Irving Lorge) to 30,000 words. His initial list of 10,000 words was compiled from 41 sources and a total of four million running words. He enlarged it to nearly ten million running words from 241 sources and 20,000 selected words in 1931. Between 1931 and 1944, he and Irving Lorge enlarged the corpus by additional studies of their own and the incorporation of prior studies by others. The resulting corpus was on the order of 23,500,000 running words from which they produced their "Teachers' Wordbook of 30,000 Words" in 1944.

One interesting feature is that like Keniston, Thorndike modified his evaluation of word importance as indicated by frequency by integrating it with its range of occurrence, i.e., the number of sources in which it appeared, designating the index the "merit number" of a word. The value of the so-called "merit number" was defined by the calculation, $MN = f/10 + r$; f being the word frequency and r, the number of sources in which the word appeared.

The range of the 30,000 list is imposing, involving 285 separately listed sources including those supporting previous counts by other researchers on reading, writing, and spelling vocabularies, the Lorge Magazine Count of one million words and a juvenile count of about 120 sources. One problem with the 1944 publication, however, is that from the point of view of research, it omits much of the background material on procedures which appeared in the 1931 "20,000 Wordbook". It also omits a list of 135 words common to all sources Thorndike used and groups the highest frequency words under gross occurrence categories ("AA" and "A") without differentiation.

The material forming the basis of these counts is now old, and much was old even in 1921, since about three of the first four million words in the original corpus came from the Bible. In spite of these drawbacks, it was much in demand and was reprinted as late as 1963. At the present time, it has been superseded for printed English for children by the American Heritage "Word Frequency Book" (1971). For the written language of adults,

13

it has been superseded by the "Computational Analysis of Present Day American English" by Kučera and Francis (1967).

The revised basic Thorndike Teachers' 20,000 Wordbook of 1931 was supplemented in 1938 by Thorndike and Lorge in "A Semantic Count of English Words" which gave the frequency of occurrence of each meaning of each word of Thorndike's 1931 Wordbook, based on a detailed analysis of some 2,350,000 words.

In 1949, Lorge published a revised and improved version of the 1938 Semantic List of the 570 most frequently occurring words.

These semantic lists helped correct a basic deficiency of those objective word lists which did not separate the frequency of the several meanings of the dictionary entry or the difference between the completely different meanings of homonyms which *are* separate dictionary entries. For that reason, the normal objective count in which pronounciation and form, not meaning, are important, has been less satisfactory from a reading standpoint, although it may well serve its purpose for shorthand and spelling.

In 1923, Godfrey Dewey, apparently dissatisfied with the lack of diversity of materials included in English frequency counts made for the purpose of teaching of shorthand, published "The Relativ Frequency of

English Speech Sounds".[1] This is the first such study of sounds recorded since Whitney's in 1874. He corrected the problem of representative selection of materials by sampling newspapers (editorials and articles), modern novels and short stories, speeches, personal and business correspondence, advertising, religion (Bible, sermons, journals and papers), science, and magazines. Dewey's error, while correcting lack of diversity, was in collecting samples that were too small. His total corpus was only 100,000 running words, which is about 1/10 what Bongers considers a minimum (1947, pages 37 and 240). It is noteworthy that oral English was included in this count in the form of speech materials.

In 1924, V. A. C. Henmon of the University of Wisconsin published his oft-cited "A French Wordbook Based on 400,000 Running Words". More than 60 people contributed to the sampling and collection process, obtaining 400,000 running words from nine different categories of printed and written sources. These 400,000 word tokens were found to represent 9,187 different word types or orthographic variants, 1,250 of these occurring 25 times or oftener. Subsequently, Henmon published a separate listing of the 3,905 words which occurred five or more times in his count. His study originated in an attempt to find the influence of Latin on French, but developed to serve much broader educational purposes, particularly vocabularly selection.

---

[1] Dewey's tendentious attempts at English spelling reform have done little more than cause secretaries and type-setters to misspell the title of his article.

In 1926, Ernest Horn, as a result of several years of study of his own and the use of the research of others in the field of analysis of personal and business correspondence, published his well known "Basic Writing Vocabulary". It was based on a total of 5,136,160 running words which provided 36,373 different words after omitting proper nouns. From these, Horn finally selected 10,000, although he found that the reliability of his count decreased rapidly after the first 1,000 with occurrence frequencies less than 77. He considered both frequency and range by use of a complicated formula to ensure due consideration of range in the final selection of words for the list. He deliberately left out all words with less than four letters since his interest was spelling and he felt that words of three letters or less are not hard to spell. He also left out 41 common words, of the type that would probably appear in anyone's list of the first 100 most common English words. They were mainly short adjectives, adverbs, and pronouns. Omission of words with less than four letters has caused problems in trying to apply the list for general vocabulary purposes. The omission of the most common words, principally functors by the time of Horn's work, was a common practice motivated by an attempt to capture a greater number of substantive words ordinarily displaced by the ubiquitously occurring functors.

Early Modern Foreign Counts

In 1927, Milton Buchanan, working under the auspices of the American and Canadian Committees on Modern Languages, produced a "Graded Spanish

Wordbook". Its principal purpose was to provide material for graded vocabulary tests. Buchanan amassed 1,200,000 running words from a total of 40 sources spread over seven categories of printed material. He believed the range was adequate for classroom purposes including oral practice in conversation, even though no actual conversations were included in the corpus. The dialogues drawn from plays were supposed to furnish the oral element. With respect to word importance, Buchanan used the Thorndike-Henmon method of combining range and frequency in order to obtain a merit number. Buchanan found 18,331 different words in his corpus, but reduced them to a list of 5,324 occurring ten times or more with a range of 1-40. Following fast-developing practice but extending it even further Buchanan also eliminated the 189 most common words from his general vocabulary, and published them separately. These deleted items consisted of articles, conjunctions, numerals, pronouns, proper and geographic names, adjectives, adverbs, prepositions and some very common nouns and verbs. Buchanan was cognizant of prior Spanish word counts such as that of Keniston but there is no indication that he incorporated them as Vander Beke did Henmon's in French.

In 1928, B. O. Morgan revised the Kaeding Frequency Dictionary of the German language, also under the auspices of the American and Canadian Committees on Modern Foreign Languages. His purpose was to make the Kaeding list useful from the standpoint of teaching foreign language, in addition to stenography. To make the count more useful for general purposes, Morgan used the concept of basic or stem words and grouped under them all words in the Kaeding count which had a cognate or semantic similarity and a frequency count of 200 or more. This grouping resulted in a list of 2,402 stem words which he arranged in blocks of descending frequency ranges. Morgan then prepared an alphabetical list of 6,000 words in which he listed the basic 2,402 words together with any of their derivatives with a frequency of 100 or more.

Although the Morgan revision made the Kaeding count more usable, it did not correct the sample problems. It was still general, printed vocabulary containing no oral sampling as such, was out of date (even in 1898) and contained no specialized words (such as those required in the classroom) either in the main list or in any supplemental list.

In 1929, Vander Beke, under the sponsorship of the American and Canadian Committees on Modern Foreign Languages, published another French word count called the "French Wordbook" which incorporated (extended and updated) Henmon's earlier work. Vander Beke's corpus amounted to 1,147,745 running words and 19,253 individual words. A cut-off at the range of five reduced the list to 6,067 words. These he made into a list using range as the main criterion rather than frequency. Vander Beke also listed the 69 most common words separately, as Part I of his study. The 69 all had a frequency on the Henmon list of 450 or more and consisted principally of structural entries.

Vander Beke set up his basic list of 6,067 in Part II in such a way as to show range and frequency of each word in his independent count, the

Henmon frequency, and a combined frequency of each word.

The combined corpus was over 1,500,000 words. (Henmon's 400,000 and Vander Beke's own 1,147,748.) Like Henmon's, Vander Beke's sources were printed or written materials, some of which went back into the mid-Nineteenth Century. In 1939, West and Bond reworked Vander Beke's list to make it more convenient for the teaching of reading by grouping derivatives under headwords and providing lists of Latin roots and French affixes to assist in word recognition. Groups of related words, were listed in frequency groups of 100.

In 1929-1930, there appeared three supplements to single word lists in Spanish (Keniston), German (Hauch), and French (Cheydleur). These supplements consisted of lists to account for fixed collocations of words which together conveyed a meaning different from the sum of the meanings of the individual words.

In 1930, C. K. Ogden published the first edition of his "Basic English". It is subjective list based on essential semantic concepts rather than the result of an objective frequency count. It is of interest, however, because, like some objective counts it contains a minimum essential or Island vocabulary. The number of words is stated as 850, but the actual count may run as high as 2,000 depending on how variants of the basic 850 are counted. The purpose of Basic English was to produce an international language. In the process, meanings and grammatical constructions of standard

English were changed so much that Basic English cannot be considered a minimum vocabulary of standard English, although its words are often compared with those resulting from objective counts of standard English. Basic English was revised and expanded by E. C. Graham in 1968. Five years earlier, Lancelot Hogben published Essential World English as his replacement for Basic English as a universal language. Instead of using an 850 word list, Hogben (1963) recommended a 1300 item list of what he calls Essential Semantic units. His list largely avoids synonyms, homonyms, and dual meanings of any unit while embracing all necessary concepts.

Early Word Counts As Vocabularies

In addition to the American and Canadian Committees on Modern Foreign Languages, one of the chief proponents of limited vocabularies of English for teaching purposes was the Institute of Research in English Teaching (IRET) sponsored by the Japanese Department of Education in Tokyo. Its head was Dr. Harold E. Palmer. Beginning in 1931, Palmer and his associates began to publish English vocabularies of 500, 600, 1,000, 2,000, and 3,000 words. As they were revised, the 1,000 word minimum word lists became the most popular. These lists introduced the idea of radius, which was almost like frequency grouping, in that each radius list contained a predetermined number of most important words; 500, for example, and the next radius might have 1,000 words which would include the first 500 plus the next most important 500. These lists tended at first to be more subjective than objective, but later became selective lists based on considerations of objective frequency and range as well as subjective and empirical considerations.

In 1934-1945, Mr. Michael West, under the sponsorship of the Carnegie Corporation, convened conferences to coordinate the efforts of the objective word counters such as Thorndike, the IRET group from Tokyo, and the teachers of English as a foreign language. As a result of two major conferences, Dr. West and his associates published the Committee report as the "Interim Report on Vocabulary Selection", in 1936. It included a list of 2,000 General Service Words to be used as a basic vocabulary of English for foreign language students. Dr. West, assisted by Dr. Lorge, revised it into a semantic frequency list based for the most part on five million running words. The list was arranged by word frequency, but with the frequency of the various meanings of each word indicated by the percentage of the frequency value of the stem word contributed by each meaning. The list contains a supplementary list of 425 popular scientific words to round out the basic 2,000 word list. Dr. West published the revised list and its supplement in 1953 as "A General Service List of English Words". In the late 1930's, West also published several other minimum vocabularies of from 900-1,500 words, generally comparable to those of the first and second thousands of the IRET 3,000 word list.

In 1937, Aibert de la Court, who was teaching Dutch in Indonesia, produced a word count in Dutch which included word combinations (idiom-like expressions). Its purpose was to assist teachers and textbook writers of

Dutch in Indonesian schools. It was a count based on 370 printed sources covering both adult and children's books, magazines, and newspapers in Java and in the Netherlands. From one million running words, 3,296 separate words and 2,000 collocations were determined. The unit of entry in the list was the head or stem word. Under it were listed derivatives and compounds with a frequency of 25 or more. Inflections were not shown on lists separately but counted under the head word. Homonyms were listed separately. Word importance was determined basically by its range. The number of derivatives and compounds of the entries were also noted. It was estimated that the two lists embraced 95 percent of the material in an adult publication in Dutch. Words which fell within the 200 most common occurrences were not included and were designated as "uncountables".

The de la Court list is a general service list. For classroom use, he added a supplemental list of 67 words as an appendix. During the 1930's, there were several attempts to improve word counts by combining them as Horn and others had done. While the corpora of the combined lists were larger than that of their component lists, the resulting lists inherited all the faults of the component lists except that of small size and restricted sampling. Beginning in 1934, Helen S. Eaton began to compare the first 6,000 words in selected lists of English (Thorndike), French (Vander Beke), German (Kaeding), and Spanish (Buchanan). Eaton started with word frequencies, then expanded them into semantic frequencies of the several meanings of the words. Her idea was to identify and correlate common concepts as expressed in the most frequently used words in the four European languages believing that if the speakers of these languages used common concepts frequently, these concepts might represent ideas basic to mankind and be found in other languages as well. Practical problems arose, however, in restricting the comparison to the semantic variations of only the first 6,000 words on each list. While a stem word may rank in the first 6,000 on a list in a given language, that does not necessarily mean that any or all of its meanings will also do so. In addition, stem words lower in rank than the first 6,000 may have individual meanings that have greater frequency than some of the meanings of the stem words in the first 6,000. In a comparison among four languages, the problems are quadrupled. For that reason, some possible correlations were not made, and some that were made on the basis of the frequency of the stem words equated very high frequency meanings of some stem words in one language with very low frequency meanings of stem words in another language. The result is that some significant concepts in one language are correlated with much less significant concepts in another language. Further, the study appears to assume that single word meanings alone represent concepts. Eaton finally published her completed work as "An English, French, German and Spanish Word Frequency Dictionary" in 1940.

Perhaps the last of the reworkings of earlier vocabulary lists through objective, subjective, and empirical means was Herman Bongers' so-called "KLM List". This list represents a comparision of several prior lists containing three thousand or more words. From these lists, Bongers derived a distilled list of three thousand words. He subdivided the 3,000 words into 1,000 word lists: K--the first 1,000; L--the second thousand; and M--the third thousand. Bongers was greatly influenced by Palmer and the KLM list is most comparable to the IRET 3,000 word list of 1932. However, Thorndike's list and a composite list by Faucett and Maki were also considered. The KLM lists are arranged alphabetically with derivatives indented under their headwords and the Thorndike frequency grouping indicated, together with Bongers' rating when it differs from Thorndike's. Inflections are not listed except in the case of irregular forms, whether plurals or verbs. However, inflected forms are considered in the frequency of the headword and separate listings are made for homonym forms When these lists were tested against ten English texts, they were found to contain 97.48 percent of all the words in those texts, with the K-list (first 1,000 words) accounting for 89.46 percent of all the words found, thus, emphasizing again the small part of total vocabulary we normally use. Bongers published his KLM list in 1947 as an appendix to a comprehensive study of vocabulary building entitled "The History and Principles of Vocabulary Control" (1947, Part III, 82 pages).

A comparable, detailed study of word lists and vocabulary, including frequency counts, entitled "English Word Lists" by Charles C. Fries and A. Aileen Traver, was first published in 1940 and republished in 1950. Both the Fries and Traver, and the Bongers' books give excellent histories and discussions of vocabularies and frequency counts up to about 1940. However, they frequently disagree on their analyses of the problems involved in the counts and in their opinions of the quality of the results obtained by individual authors. Overall, Fries and Traver are more general, and more inclined to description than criticism.

In the early 1930's, under the influence of the IRET a number of Japanese investigators attempted to identify a minimum basic vocabulary. Most of the vocabularies so conceived were subjective and/or empirical, and contained from 1,000-2,000 words. Since 1950, however, the Japanese frequency counts, especially those conducted under the sponsorship of the National Language Research Institute (NLRI) (or Kokuritsu Kokugo Kenkyugo) have shown considerable sophistication in vocabulary building by statistical methods. The following three Japanese language studies are representative.

Modern Foreign Language Counts

Japanese

In the early 1950'[illegible] the NLRI started on its "Research in Modern Vocabulary" which investigated the vocabulary used in women's magazines and cultural reviews. Part 1, published in 1953 (and often cited as a separate study), gave the report on the research based on sampling the text of one year's issues of two women's magazines which were considered representative of that type of publication. A corpus of three million running words was compiled. Part II, published in 1958, gave a report on research based on a sampling of 13 cultural reviews which resulted in a corpus of 230,000 running words.

The analyses made of the findings in each case considered mainly the statistical and semantic structures of vocabulary and word construction. Much use was made of statistical sampling as opposed to the more laborious word counting done by Thorndike and Horn and their associates. Each of the parts contains a listing of the 4,000 most frequently used words. (The National Language Institute of Japan, 1953 and 1958).

In the late 1950's, the NLRI undertook another study of vocabulary and Chinese characters found in modern magazines. It covered the fields of culture, business, popular science, housekeeping, sports, and other amusements. The NLRI published the results in three volumes in 1962-1964. Volume I (1962) was entitled "General Description of the Project and Vocabulary Frequency Tables". Samples included 540,000 words out of a total corpus of 1,400,000 words. From the 540,000 words the 7,200 most frequent were published in various forms in a series of eight tables. Volume II (1963) contained the "Chinese Character Frequency Tables", giving not only the 1,995 most frequently used Chinese characters but also the total 3,328 Chinese characters officially used in Japanese. Volume III (1964) is called "Analysis of Results". However, it also gives much data not given in the first two volumes in addition to the details of the procedures followed. For example, it gives the 1,200 most frequently used words with semantic classifications of the first 700 of them, the statistical structures of vocabulary, bound forms (idiomatic expressions), compound words, and homonyms.

Both of the above Japanese studies were based on the printed material found in periodicals. The only study of comparable scope which has come to our attention is one by H. Miyaji published in 1971. Miyaji built up a frequency dictionary from a 250,000 word sampling of Japanese fiction, periodicals, drama, didactic prose, and scientific writing. Its full title is "A Frequency Dictionary of Japanese Words".

Russian

The first major count of Russian was published in 1953 by Harry Josselyn. It was basically a computerized analysis of literary Russian of the period 1825-1950. Its purpose was to determine word frequencies and frequency occurrence of categories of Standard Literary Russian.

The percentages of the total material collected were 25 percent from the period 1825-1899, 25 percent from the period 1900-1918, and 50 percent from the period 1919-1951. The styles range from drama, 7 percent, to fiction, 59 percent. Oral language is included indirectly since samples were selected to contain 37 percent literary conversation. The purpose of the count was to assist in the teaching of Russian as a second language. In common with recent practice, the count contains a list of the 204 words most likely to occur in all similar counts of Russian. These words are not included in the count proper.

In all, one million running words were recorded of which 526,044 were actually used. Of these, 41,115 were different words. From the 41,115, a list of 5,230 significant words was published in four lists of approximately 500 words each and a final list of the remaining 3,000. The first 490 words were broken down by range, frequency, time period (in which written), type of literature, and categorized as conversation or non-conversation. The remaining words were listed in rank order determined largely by range. This list can hardly be called current or colloquial, but may be of assistance in developing courses of instruction for personnel who wish to read Russian.

The second Russian word count is that of N. P. Vakar. Significantly, it is called, "Spoken Word Count". It is divided into two parts: Volume I, Vocabulary (1966), and Volume II, Sentence Structure (1969).

In view of Soviet Russians' reluctance to talk into foreign tape recorders, (for Part I) Dr. Vakar resorted to an indirect method similar to, but more extensive than, Dr. Josselyn's. Vakar took 50-word samples from each of 200 acts of 93 plays, published between 1957 and 1964 to ensure currency. These samples provided a 10,000 word corpus which is small by most standards, but which Dr. Vakar believed to be sufficient for colloquial oral Soviet Russian. He found 2,380 different words in the 10,000 word corpus. He also found that 360 of the 2,380 words account for 73 percent of the words used in Russian conversation as represented by the samples.

In Part II, Sentence Structure, Vakar analyzed the material in terms of "kernel" sentences. Some 1,000 sentences were selected for analyses from a statistical universe of one million running words found in the same plays which were sampled for vocabulary in Part I. One of the findings is that spoken colloquial Russian varies considerably from literary Russian and that short sentences of 1-5 words make up 75 percent of the total utterances in oral Russian.

If we can assume, as the author does, that modern Russian drama is a true representation of colloquial Russian speech, this is an excellent statistical study of current-day oral Russian. The author validated his study by comparing its findings with those of several other Russian word counts including Josselyn's and noted the differences and similarities.

Spanish

In Spanish, two word counts made in the early 1950's deserve note. The first was done at the University of Puerto Rico by the Superior Teaching Council of Puerto Rico under the directorship of Dr. Ismael Rodriguez

Bou with Dr. Lorge acting as consultant. The word count is called Spanish Vocabulary Count (Recuento de Vocabularie Espanol). It is a modern computer compiled frequency count published in 1952 to provide for Spanish the teaching materials already existing in English through the efforts of Thorndike, West, and others.

This is a comprehensive word count embracing both active and recognition vocabularies, written, printed, and oral materials, and both adult and children's vocabularies. The total corpus is 7,066,637 running words, including Buchanan's corpus and that of an unpublished Spanish count made in Puerto Rico by two members of the Faculty of the University of Puerto Rico: Dr. T. Casanova and A. Rodriguez, Jr.

About half of the running words were active (speaking/writing) words, and the other half recognition (reading/listening) words. The active vocabulary (about 3,390,000 words) was made up of children's oral, written (including the Casanova/Rodriguez input) and association inputs. The recognition vocabulary (about three million words) came from periodicals, radio programs, religious materials and the Buchanan corpus. In addition, there were about 700,000 words chosen subjectively by the authors from school texts and supplemental reading materials.

The count also contains the results of the analyses of children's conversations and their association vocabulary. The children's material throughout was from elementary grades 1-6 except for the Casanova-Rodriguez corpus which was taken from compositions written by children in grades 2-8. The oral vocabulary was taken down stenographically or recorded electronically. Great care was taken to obtain samples representative of Puerto Rico geographically and of children in all phases of their daily school life. For the oral samples, great care was taken to do the recording unobtrusively so that it would represent spontaneous conversation. The "association" vocabulary was of two types: "controlled" and "free". Controlled association responses were evoked by stimulus words selected from a prepared list. The children were told to write all the words which occurred to them after the stimulus word was spoken. Free association lists were produced by asking school children to write down all words occurring to them in five minutes.

Neologisms and regionalisms were included in the corpus as were "coined" words not in standard dictionaries, if judged to be common among educated people.

Frequency was the criterion for rank order of words in the lists. Inflectional forms were included but semantic frequencies were not.

The seven million running words resolved themselves into 83,430 different units: 20,542 lexical and 62,999 inflectional forms. Part I of the count deals primarily with an explanation of the count and presentation of the first 10,000 lexical and first 20,000 inflectional units listed in order of frequency and alphabetically. Part II contains all lexical units and their inflectional forms classified by total frequency and frequency of appearance in various texts. Excluded from the count, but listed in an appendix are the 105 most frequent words of the count.

The second count in Spanish is that of Victor Garcia-Hoz: "Usual, Common, and Fundamental Vocabulary", published in Madrid in 1953. Garcia-Hoz also distinguishes between active vocabulary (speaking/writing) and latent (or recognition) vocabulary (listening/reading). However, he uses as the source of his corpus only four major categories of materials. He took a 100,000 word sample from sources in each category for a total of 400,000 running words. The categories and sources or materials were as follows.

Aspect of Living	Category of Material
Private or family life	Private letters
Unregulated social life	Periodicals (Newspapers)
Organized social life	Official documents of government, church, and labor unions
Cultural life	Books and reviews.

This is an adult word count. It can be considered to include oral material only in the sense that private letters are part of active vocabulary and that words written may also be customarily spoken by the writer.

The 500,000 running words are distributed in descending order of quantity and ascending order of frequency in such a way that the usual vocabulary includes the common and fundamental vocabularies and the common vocabulary includes the fundamental. Significant data on the lists appear below:

Vocabulary	Description	Words	Average Frequency
Usual	Language of the common man	12,911	31 (4.2)
Common	Frequency between 40-399 and appears in all four cate- ories	1971 plus Supplemental list of 212	172 (52)
Fundamental	High frequency (**400-up**) are evenly distributed among all four categories	208	1324 (1324)

Looking at frequency of the categories in another way, if we take the common vocabulary (which includes the fundamental) out of the usual, the average frequency of the remaining words is 4.2 or about one per category on the average. If we take the fundamental out of the common vocabulary, the remaining words have an average frequency of 52 or 13 per category on the average. By itself, the average frequency of the fundamental vocabulary is 1324 or about 331 per category of material on the average. Thus, the fundamental words are truly the commonest words in frequency and range. Words with high frequency (over 400) but of uneven distribution were not included in the fundamental list. 26 words were left out for this reason; 19 had too high a frequency in writing and seven had too low a frequency in writing. In determining the common and fundamental categories, the author made extensive use of mathematical techniques, including factorial analysis, to establish correlations among the four categories of material and the three types of vocabulary.

As a test, the author compared the words in the vocabulary with the language used in Spanish drama, to determine whether the words were colloquial and current. In this, he agrees with Vakar that drama contains most of the colloquial language of its time. In the normal "periodical" category, the author omitted sampling magazines on the basis that they are hybrids between newspapers and books and their words would be included already.

This vocabulary analysis, like that of the University of Puerto Rico, does not extend itself to semantic frequencies, nor does it really involve oral language. However, this count is noteworthy for its ordering of telescoping vocabularies and for its mathematical computations of the correlations underlying the selection of words for inclusion in the common and fundamental vocabulary.

French

In French, there has been one recent frequency count of special interest. It was prepared by the National Pedagogical Institute for the French Ministry of National Education, from 1954-1964. It is called Fundamental French and consists of First and Second Level (Stages) and an Elaboration on the First Level. Fundamental French (1st Level) (French Ministry of National Education, 1959), replaced "Elementary French" which appeared in 1954 in response to Basic and Universal World English without the restrictions on growth inherent in the Island Vocabulary of Basic English. The 2nd Level appeared in 1962, to provide vocabulary and grammar for teachers of students who wanted to extend their knowledge of French beyond the necessities of daily life. The elaboration of the First Level (Goughenheim, et al., 1964), provided the detailed background and procedures leading up to the First Level.

The purpose of Fundamental French was to provide vocabulary and grammar for teachers instruction foreign students. The first level was fundamentally spoken or oral French, based on an objective and a statistical approach. There are some discrepancies between the explanations given in the report on the First Level and that in the Elaboration, but the general procedures and results are given below.

Informants recorded their conversations on tape recorders as spontaneously as possible under the guidance of research assistants. Informants from all over France were interviewed. There was an effort to cover as great a variety of professions and vocations and as wide a range of subject matter as possible to obtain representative samplings. The 275 informants were mainly adults, about evenly divided among men and women, but also included 11 children of school age. There was also a good spread of educational backgrounds among informants with perhaps the greatest percentage

(37 percent) having completed formal education through the primary grades, only. In all, a corpus of 312,135 running words was compiled yielding 7,995 different words. The frequencies varied from 14,083 to 1, and the range from 163 to 1 (the material of the 275 informants had been combined into 163 units for examination). For the purpose of the First Level (Basic), the lexical list was selected from words with a frequency of 29 or above. This provided a lexical list of 1,963 words. It was a frequency based list with range considered only to differentiate among words of the same frequency. When both frequency and range of two words were the same, the words were listed alphabetically. In the final list, the lexical units were arranged alphabetically, with no indication of their frequency, since as far as teachers were concerned they were all equally important. In common with most counts we have observed, the most frequent words were the grammatical or structural ones. In the French count, interspersed at lower frequencies in order of first appearance, were verbs, adjectives, and nouns.

As would be expected from the above, it was determined by comparison with written counts that certain very useful words, particularly nouns, but also verbs and adjectives have only low frequency in written or oral counts taken from general or random samplings. These concrete words applicable to specific situations and subjects get crowded out of frequency counts by the general usage words, of which the grammatical words are the prototype. The authors called these concrete words (which are needed even in a basic vocabulary but appear in general word counts with only very low frequencies, if at all), "available words". Everyone has to know them, but the occasion for their use occurs only infrequently. In this way, "availability" becomes a second principle of Fundamental French along with frequency. To determine what "available" words to use, the researchers resorted to a controlled association type of collection covering 16 interest areas, such as "furniture", using 904 elementary school students aged 12-13 of the Departments of France. Each student supplied 20 words per subject area. Those of highest frequency were added to the First Level vocabulary.

Although a semantic frequency count was not made, the words on the list were checked for meaning and where concepts essential for educational or communicative purposes were missing, words to convey them were added. This procedure added about 400 words. The list was then culled to eliminate certain words which, although warranted by frequency, were close synonyms of words of higher frequency, were vulgar words, difficult to learn, or for some other reason failed to conform to the objectives of Fundamental French.

The final list of the First Level contained 1,445 items; 1,176 lexical words and 269 grammatical words. The grammatical words chosen were the minimum deemed required to permit flexibility in the use of the language. The lexical list had a general alphabetical list of all words, followed by special lists of related words such as numbers, days of the week, months of the year, and seasons. The list was kept deliberately general with the exception of the items indicated above. It was designed to be a minimum vocabulary to which specific additions could and would be made by teachers according to the environmental needs of their pupils, especially regionalisms to adapt the standard language to the needs of particular geographical areas of France.

For teachers of those who wished to go beyond the ability to express the daily needs of life and to acquire a more complete knowledge of French and French culture, Fundamental French (Second Level) was developed. Unlike the First Level which is largely based on the oral frequency count, the Second Level is based on the written language and includes additional grammatical terms, in order to provide the student considerable flexibility of expression and an ability to read newspapers and books.

The First Level took in words from the original word count down as far as frequency 29. The Second Level lowered the threshhold to 20 or above, and included many of those above 29 which had been rejected as not required for the First Level, particularly those which were eliminated by reason of duplication of basic concepts. The Second Level also adopted many of the terms on the association lists of the 16 interest areas which had not been deemed to have sufficient frequency to warrant inclusion on the First Level. In addition, the authors took words from the Vander Beke list with a frequency of 60 or more, even though that list was both literary and dated. Next, they undertook new investigations and short counts to update Vander Beke's count. One field was newspapers and magazines. The researchers counted words appearing under 14 subject areas in the newspapers and magazines and added an average of 35 words from each of the subject areas if not already in the First Level list as amended. These additions amounted to 425 words with a frequency of 13 or higher. Further, using the association method, 160 students at teacher preparation institutions throughout France furnished lists of psychological terms. Those used by 15 or more of the informants were added to the list. Finally, the list was submitted to a panel of experts who added such words as deemed by them to be required to meet the purposes of secondary level French instruction. Like the First Level, the Second Level of Fundamental French contains an alphabetical list of lexical units, and a section of grammatical words.

Note that in Fundamental French, the vocabulary lists are a combination of objective frequency counts, empirical inclusion of concrete words, exclusion of duplicating words and those of low frequency, and inclusion of other words based on empirical association by students, and an addition of still others based on the subjective judgment of panels of experts.

Fundamental French is of interest not only because the first level is oral but because it provided a point of departure for Dr. J. Alan Pfeffer of the University of Pittsburgh in a study of oral German which will be discussed next.

German

In German, there have been three recent studies--one general and oral by Pfeffer, one on newspaper vocabulary by Rodney Swenson of Hamline

University in St. Paul, Minnesota, and one by Scherer of the University of Colorado on the "Short Story in the Second Quarter of the 20th Century". Dr. Pfeffer's is the more representative of the three. Since it is basically oral, it also is best suited to our purposes. For that reason, it will be described briefly. In many ways, it is one of the best of the modern word counts, having profited from the faults of prior studies. So far, of eleven expected publications to result from his study, Pfeffer has published three: Basic (Spoken) German Word List (1964), Index of English Equivalents for the Basic (Spoken) German Word List (1965), and Basic (Spoken) German Idiom List (1968). Before undertaking his study, Pfeffer reviewed the field of word counts and noted the best features of the recent ones, especially the Spanish Word Count produced by the University of Puerto Rico (Rodriquez Bou, 1952), and Fundamental French, produced under the auspices of the French Ministry of National **Education (1959,1962). In** general, Pfeffer appears to have followed, but improved upon, the procedures used by the authors of Fundamental French (First Level) and provided oral, topical (utility or available) and empirical inputs to his own corpus of oral German.

The first step was the collection of the oral vocabulary. This was done by means of taped interviews on 25 human interest subjects. The interviews took place in 56 cities and towns in Austria, German-speaking cantons of Switzerland, and West Germany. Basic data such as age, sex, educational background, vocation, and type and size of residence were recorded for each informant. 401, 12-minute recordings were transcribed and the words placed in context on ADP punch cards. In this process, proper names, place names, and faults of speech were deleted as being peculiar to the individual or place. In this way, an oral corpus of 595,000 lexical units was derived from which nearly 25,000 separate lexical units were isolated. Inflections were subsumed under their headwords, but their frequencies separately recorded. The frequency varied from 50,256 to 1 and range (of speakers) from 450 to 1. (Some interviewers' conversations were also included, so the 401 interviews developed into a range of 450 speakers.) From the 25,000 separate lexical units, nearly 1,000 representing the most common words with frequency at least equal to 40 and range equalling at least 25 were selected for further analysis. The analysis was concerned mainly with applicability, universality, and indispensability. This screening process reduced the list from 1,000 to 737 spoken words. (The oral part of the corpus.)

The utility (topical) words were collected by controlled association in 82 intermediate and academic high schools in 48 German, Swiss (German speaking) and Austrian cities and towns. The informants were about 15-16 years old of both urban and rural backgrounds, and about equally divided as to sex. The students were given a stimulus topic selected from a list of 21 such subjects, such as "buying and selling". They were then given ten minutes to write down 20 nouns (or 12 verbs and eight adjectives) related to the stimulus topic. (Whether nouns or verbs and adjectives were to be collected and on what topics was specified in the request to each school.) The effort yielded a topical corpus of 833,000 terms from which

19,700 nouns, 6,800 verbs and 7,400 adjectives were derived. Applying the the criterion of applicability to the topical list narrowed it to 347 nouns, verbs, and adjectives.

In the empirical stage, the 737 oral words were combined with the 347 topical words and all of them were examined together for gaps in sequence, derivation, opposites, topical limitations, parts of common compounds, and common concepts. The result was the addition of 185 words to round out the basic list. About three-fourths of the words had already been considered in the uncut oral or topical corpora, but had been eliminated, generally because their range, frequency, or both had been too low. The resultant total word list consisted of 1,267 words. They were presented in alphabetical order (by family groups), then by parts of speech, and finally in order of frequency and range.

The index of English Equivalents (1965) gives the most common 75 percent of semantic meanings, and indicates the percentage of the headword represented by the frequency of each meaning listed. From this list, teachers can easily determine which of the several current meanings in oral usage are of most importance for students to learn. For background on handling semantics, Dr. Pfeffer leaned heavily on Dr. Lorge's treatment in his semantic analysis of the 570 most common English words published in 1949.

In the idiom list, Pfeffer defined an idiom as a "semantic restriction of syntactically collocated parts". In the gathering of the corpus of 595,000 running words for the basic oral vocabulary, Pfeffer identified nearly 25,000 single words and 7,500 phrases. From the 7,500 phrases, he extracted 1,026 idioms, an additional 99 idioms from the utility and empirical studies. The total of his idioms is thus 1,125. In his study of idioms, Pfeffer compared his list with that of Hauch published in 1929, and indicated which of the items in his list were also in Hauch's.

Dr. Pfeffer estimates that his Basic Word List, Semantic Equivalents and Idioms account for about 85 percent of the free forms, and of the restricted forms and patterns, in colloquial German speech of the present day.

Swahili

In Swahili, there appears to have been no major or comprehenisve frequency count of the written or oral language. There have been subjective and empirical studies made which resulted in grammars and dictionaries, however. Bilingual dictionaries, for example, have appeared in several European languages: Swahili-English (French, German, Polish, and Russian). Missionaries started compiling grammars and vocabularies, which grew into dictionaries, as early as the 1850's. Dr. Krapf published his dictionary in 1882 followed by Madan in 1903. Perhaps the best known dictionary in English was compiled by Frederick Johnson in 1939. In the same year, a well-known French-Swahili dictionary compiled by Charles Sacleux appeared.

One of the latest in English was published by D. V. Perrott in 1970. With regard to grammars, one of the first in Swahili appeared in 1850, prepared by the same Dr. Krapf who later published one of the earlier dictionaries. His grammar was shortly followed by Bishop Steere's in 1870. Most of the grammars contain vocabularies for each lesson and a glossary of all words used as an appendix. While many are more interested in translation and writing than conversation, there is an increasing number which devote considerable space to conversation, as exemplified by the publications of the Foreign Service Institute of the Department of State which includes "Swahili--an Active Introduction (Conversation)" (Stevick, et al., 1967). Other good grammars are Edgar Polome's "Swahili Handbook" (1967), and D. V. Perrott's "Teach Yourself Swahili" (1951, 1967). The Belgians have also been interested in Swahili because of their interest in the Congo, particularly in Katanga where a dialect form of Swahili called KiNgwana is spoken. In the 1940's, Van den Eynde developed his "Grammaire Swahili", (1944), but considered the Katangan dialect so bad he concentrated on the so-called Standard Swahili of the East Coast. On the other hand, E. Natalis in a three volume work called "La Langue Swahili" which appeared in 1965, addressed principally the dialect of Swahili spoken in Katanga.

In recent years, there have also been some specialized studies, principally by students and scholars on the various aspects of Swahili Grammar. In the United States, there seems to have been a concentration on the verb. Carol Eastman made a study of verbal extensions (1967), Carol

Scotten delved into the extended verb system (1967), and Rae Moore made a study of verbal derivations (1966). On the other hand, Judith Olinick became interested in exploring transformational grammar as it relates to certain noun phrases (1967). In spite of these recent studies, many of them done with the aid of tape recorders and computer manipulation of results, there is still a need for an extensive frequency count in this language, similar to the latest ones done in the European languages and Japanese.

Comparative Studies

In the field of comparative linguistics, Kucera and Monroe (1968) published "A Comparative Quantitative Phonology of Russian, Czech, and German". This study attempted by comparative analyses to determine the value of a statistical approach to historical phonology by studying the differences and similarities in historically related or geographically contiguous languages. The study was based on the printed word, principally prose fiction (60 percent) with half the rest of the words taken from periodicals. As a result of their study, the authors concluded that a close genetic relationship of two languages (e.g., Russian and Czech) is likely to show up at the phonological level in similar phonotactics but not necessarily in similar phonemic systems. On the other hand, languages in close geographic contact (e.g., Czech and German), may well show the greatest similarity at the phonological level in phonemic inventory, with much less similarity in their phonotactics.

Modern English Studies

Structural Analyses

In the field of structural analysis, the first study was on traditional frequencies of English phonemes by Hultzen, Allen, and Miron (1964). The corpus for their study was developed by Professor Agard of Cornell for some of Dr. Carroll's studies. It consisted of material from 11 plays and selections from the Journal of Modern English. From these sources, some 20,000 phonemes were collected in phoneme sequences. The phoneme corpus was manipulated by computer to produce displays with supporting tables of the number of occurrences (1) of each phoneme, (2) of each two phoneme sequence, (3) of each three phoneme sequence, and (4) of each four phoneme sequence.

The second study of structural English by A. Hood Roberts extended the Hultzen, Allen and Miron counts by making a quantitative analysis of the segmental phonemes contained in Horn's "A Basic Writing Vocabulary of 10,000 Words" (1926) and Lorge and Thorndike's "A Semantic Count of English Words" (1938) supplemented by Lorge's "Semantic Count of the 570 Most Common English Words" (1949). The Horn vocabulary items were spoken lists in sentence patterns and recorded on tape in north central dialect. The 10,000 words were transcribed phonemically and their etymologies tabulated. The results were then manipulated by computer and analyzed to produce tables listing the frequency of occurrence of the phonemes, average word length in phonemes, transitional probabilities of phonemes, and the etymological composition of English according to proximate sources (e.g., French as compared to the more remote Latin).

Adult Oral Word Counts

There have been six important frequency counts of oral English since 1950. Two are of children's speech, two of college students and two largely of the general public.

The first, in 1955, was that done by Black and Ausherman of the speech of students in classroom situations. Actually, the college students were servicemen of college age and background who were taking college courses in preparation for becoming military meteorologists. The informants were 274 male students who participated in 607 five-minute classroom speeches of which three and one half to four minutes of each were recorded. The students were unaware of the recording. The students were actually giving nearly extemporaneous speeches on material connected with meteorology or its background subjects, and related to its military application. The students had prepared outlines of the topics to be covered in their talks. but otherwise the speeches could be considered spontaneous.

The informants as a group were mid-westerners, highly intelligent, had good prior scholastic credentials, and high aptitudes in mathematics.

As a group, they might be considered atypical on the side of high aptitude in mathematics and high degree of prior academic achievement.

A corpus of 285,152 running and 6,826 different words was compiled with frequencies ranging from 15,000 to 1. Comparision with the Thorndike Teachers' Wordbook (1944) (printed English) showed differences and inconsistencies. There were many words in the Thorndike list which were not in the oral list, and vice versa. The discrepancy amounted to about ten percent of the oral list. Thorndike's first 1,000 words accounted for only 140 (14 percent) of the first 1,000 words of the oral list. Comparisions were closer in the case of Godfrey Dewey's Relativ Frequency of English Speech Sounds (1923). Dewey's first nine words making up 15 percent of words used amounted to 22 percent of the oral list. All the first 50 most common oral words were found in the first 83 of the Dewey list, and all but three of Dewey's first 50 were found in the first 100 oral words.

These comparisons with the Thorndike and Dewey lists are not entirely appropriate since the two printed counts are considerably dated. Other differences were introduced by the fact that the informants tended to neologisms, slang, occupational jargon, and colloquial compounds largely related to their prospective work in the military and the cultural subarea in which they were raised.

The second so-called adult oral word count was conducted by Davis Howes in the Boston, Massachusetts area during the period 1960-1965 (1966).

Informants were 20 sophomores at MIT and Northeastern University and 21 patients at the Veterans' Administration Hospital in the Boston area. The 21 patients were free of cerebral defects and any debilitating disease. 40 of the informants were taped in the course of free speech delivered in response to general questions designed to produce natural and spontaneous speech. This procedure was kept up until 50,000 words had been obtained from each informant. The 41st informant provided ten of the total 50 interviews in order to give data on the stability of word frequency. The total corpus was 250,000 words which were cataloged by source and origin, i.e., school or VA. 9,699 individual words were identified, but 4,097 (47 percent) of them occurred only once.

The study confirmed findings of others that oral language uses fewer words (has a lower type/token ratio) than printed/written English and that only very large counts of running words would reveal very rare words. In contrast to most counts, popular and place names were recorded and counted as well as certain utterances which were non-words and/or markers (e.g., mm, uh, etc.).

Howes undertook the count to update prior counts and correct deficiencies in them; i.e., Thorndike lacked an oral input and the Bell Telephone count of 1930 (French, Carter, and Koenig) collected speech sounds useful for technical purposes but in a manner not likely to provide assistance in a count of normal spoken vocabulary.

The third "adult" oral word count was published by Lyle V. Jones and Joseph M. Wepman in 1966. This count samples the utterances of 54 adults. They ranged in age from 18-80, but were mostly in the older half of the bracket. Educational backgrounds of the individuals in the group varied from 2nd grade to PhD. 20 picture cards from Murray's Thematic Apperception Test of 1943 were used to stimulate spontaneous conversation. The mean number of words per subject thus evoked was 2,527, with a range of 1,032 to 5,276. The total corpus was 136,450 running words. The results were tabulated and manipulated by computer to provide three lists:

A. The 1,102 words most often used by the 54 speakers, down to a frequency of 4/100,000.

B. Words with a range of at least 2, arranged by grammatical class and alphabetically within class.

C. List B in straight alphabetical order including inflectional forms.

The results showed little difference in word diversity between male and female or between those over and under 60, but distinct differences among socio-economic-educational groups.

This limited study indicates that 33 words account for 50 percent of the oral words used. This is half as many as estimated for the written and printed languages and generally confirms earlier studies in this respect. Jones and Wepman attribute this lesser diversity of oral speech and tendency to repeat frequent words more often in talking than writing to the fact that meaning is conveyed in face-to-face contact by bodily movements, facial expressions, eye contact, and intonation, whereas additional words are required in writing to ensure that the intended ideas are, in fact, conveyed.

Zipf's Law of the inverse relationship of word length to frequency was borne out by this study up to a word length of four letters, i.e., down the frequency list to about the 100th word in order of descending frequency rank.

The fourth "adult" oral count by Kenneth Berger in 1967 is entitled "The Most Common Words Used in Conversation". It is mentioned because of the "conversational" aspect and its clandestine (perhaps unethical) method of the collection of its corpus. Others have despaired of obtaining really spontaneous speech, e.g., the field workers for Fundamental French (1959) and Pfeffer in his collection for Basic (Spoken) German (1964). As a result, most spoken speech samples, until Berger's count, have to some extent lacked complete spontaneity. However, Berger was able to obtain unguarded conversations from bars and restaurants. His unwitting informants were largely white, male, businessmen, white collar workers, and skilled laborers. There seemed to be few professional, farm, unskilled workers or students involved. The speech collected is that of Kent, Ohio and its vicinity. Berger deveeloped his own criteria for acceptance of utterances which make his study somewhat different in methodology as well as subject matter. He accepted as sentences utterances of as few as two words which had a predicate or was a complete, although laconic, answer to a query. Slang, curse words, mispronunciations, and ungrammatical expressions were accepted. No more than four sentences from any conversational group were accepted to ensure variety within the small corpus (25,000 running words). Words normally eliminated, such as family name, place names, and other specific nouns, are listed in appendices. Berger tabulated variants under the stem word unless the forms or variants added a syllable. Forms or variants with a different number of syllables were given separate listings if the variant and its stem word each had a frequency of more than one, and if the variant and its stem word were both used with about the same frequency. The number of sentences transcribed was 2,418, with a mean sentence length of 6.7 words, representing 2,507 different words. Almost half of the 2,507 words appeared only once. Significant findings included: (1) frequent use of "I" and "you", (2) use of indefinite and relative pronouns in lieu of nouns, (3) simplicity of language, and (4) confirmation of Zipf's five generalizations regarding inverse ratios of word length and frequency, and the number of words used and frequency. Speculative findings are that conversational speech vocabulary is extremely sensitive to place, time, and current events and is subject to rapid evolutionary change.

Children's Oral Word Counts

The Beier, Starkweather, and Miller (1967) study was undertaken to determine the psychological parameters governing children's communications and also to determine whether Zipf's Laws as derived from printed/written counts were applicable to spoken counts of children's language.

The experiment took place in grades 6 (age 12) and 10 (age 16) in the Salt Lake City Public Schools. The 30 informants were all boys; half in each grade. They were selected to have a normal IQ range (90-110). In addition, data on their scholastic performance was obtained and recorded. The stimulus material is not stated in the report, but each boy recorded about 5,000 words from which about 2,700 were selected and compiled into two 40,000 word corpora (one for each grade) for a grand total of 80,000 words. Five one-minute samples of each boy's contribution were timed to obtain a rate of speaking for each informant. Comparisons were made between age groups and with the Eldridge frequency count of newspaper English in the Buffalo area in 1911.

The results tended to confirm prior findings of greater variety of expression in printed language than in speech. However, the validity of the results may be undermined by the fact that adult newspaper English of 1911 in the Buffalo area was compared with the oral language of school children in the Salt Lake City area in 1966. It should be expected that adult oral conversation would show greater diversity and variation than that of children of the ages used in this study. It would, therefore, have been better to have compared this count of children's oral English with that of a printed count of about the same date. In any event, the findings confirmed that for those two age groups and the small corpus obtained, Zipf's Law applied to oral as well as printed language. Specifically, the number of words of a given frequency increases as the frequency of uses decreases, and the shorter the word the more frequent its occurrence.

The study also determined that eight of the first ten words on the 6th and 10th grade lists were the same, although not in the same order. Other findings indicated that the 16 year olds, as compared to the 12 year olds spoke faster and used significantly more positive and negative words, slightly more singular self-reference words, slightly fewer plural self-reference words, more "other" references, and slightly more "question" words. At equivalent intelligence levels, age made little difference in the ratio of different to total words.

The second spoken word count of children's language was that of Wepman and Hass published in 1969. The children in this study were of ages 5-7. The count was undertaken to update and extend prior counts of the oral language of children in order to obtain information on grammatical development, semantic extension, and vocabulary increase as correllated with chronological age. The informants were 90 children (45 male and 45 female) equally divided among ages five, six, and seven. They were all from middle income homes and large urban areas well distributed around the United States. All were uni-lingual English speakers and had no apparent mental or physical handicaps.

The Murray Thematic Apperception Test of 1943 was used, with each child asked to tell stories about 20 picture cards. The material was then manipulated by computer.

The results were arranged in three lists for each age group, as follows:

A. Word frequency order (for all words with a range of at least two) of stem words.

B. Words by grammatical class, alphabetically within class. If a word was used as two parts of speech it was listed under each.

C. Alphabetical--including all inflectional forms and grammatical uses.

The report states no conclusion, but introduces two new concepts--a "mean" frequency for each age group on the basis of 10,000 words and all 30 informants, and a "variation" which represents the difference in the frequency of use of the word by high and low users as compared to the total number of users. A high variance index indicates that some children use the word frequently and some very little, and is, therefore, another index of what other researchers have called range. It is useful in comparing words of equal mean frequency of use since it permits estimating whether the mean frequency represents general use or use by only a few.

Printed Word Counts

In the field of printed counts, two good reports have appeared recently; one by Kucera and Francis on adult language and one by Carroll, Davies, and Richman done for American Heritage on children's language.

In 1967, Henry Kucera and W. Francis published "A Computational Analysis of Present Day American English". It essentially replaces the Lorge and Thorndike "Teachers' Word List of 30,000 Words" published in 1944.

A corpus of nearly one million words was compiled from recent and current publications, dating from 1961. To ensure adequate coverage, 15 categories of material were included: newspapers (editorials, and reviews); religion, skills and hobbies, popular lore, literature and biography, government documents, learned and scientific, fiction (five--general, mystery/detective, science, adventure/western, and romance/love story), and humor.

500 samples of 2,000 words each of continuous discourse were randomly selected for transcription and computer analysis. The results of the analysis were displayed in two ways: word lists and statistical tables and graphs.

The word lists are principally of three types: (1) descending order of frequency, (2) alphabetical, and (3) the first 100 most frequent words by total frequency and by frequency in each of the 15 categories of sampled materials. The statistical tables tabulate both word frequency distribution and sentence length distribution.

The grammatical (sentence length) analysis with frequency distribution is an added dimension to English Word Counts. For the samples as a whole, the sentences had a word length of 19.24 words[3] with sentence length ranging from 25.49 words for governmental documents to 12.76 for fiction and mystery stories.

Although this is an excellent example of a current objective word count, it would have been better had it included a semantic frequency count as well as a lexical count.

In 1971, John Carroll, Peter Davies, and Barry Richman completed a word count of the printed language of children. It was published by the American Heritage Publishing Company and the Houghton-Mifflin Company as "Word Frequency Book". Although it is useful for many educational purposes, it is primarily intended--like West's Definition Vocabulary of 1935--as the basis for a dictionary; this time a revision of the American Heritage School Dictionary. It is based on the printed language to which public and parochial school children in the United States are exposed in grades three through nine. The samples were taken from publications covering 22 subject areas: 17 curriculum areas, three library categories, magazines and religion. The curriculum categories alone sample 1,045 items (texts and other published materials) recommended by nearly half of the schools which responded to a questionnaire concerning published materials used by students in 1969.

The words to be analyzed were taken in 200 word samples from the selected printed materials until a total of 5,088,721 tokens had been amassed. The types in this corpus were determined to be 86,741. The words, after computer processing, were displayed in two types of output, only the second of which appears in the Word Frequency Book. These were: citations--occurrences of types extracted in sufficient context to provide for analysis for definitional purposes--and descriptive statistics--frequency of occurrence and distribution.

---

[3] Compare with Dewey's (1923) finding of 19.6.

The Herdan/Carroll lognormal model was used for computations. Results are tabulated alphabetically indicating total frequency, frequency of occurrence by grade level and subject, and an index of distribution (range). Unlike many other objective frequency counts, this book includes proper names, place names, and numbers. Results are also tabulated in frequency rank lists and freqeuncy grouped distribution lists, by total, grade level and category of material.

This is an excellent current frequency count of the printed vocabulary to which primary grade and junior high school children are exposed. Its source material is wide and representative and its corpus ample (five million words). It would have been more helpful to the teacher if a semantic frequency count had been included. However, the material on which such a count could have been made is available and is being used in the revision of the American Heritage School Dictionary. Hopefully, a semantic count will follow.

Summary

The science of objective or direct word counting has come a long way since the Kaeding Count of 1898. Oral counts such as those of Pfeffer (1964, 1965, and 1968) on Basic Spoken German and Jones and Wepman on U.S. English (1966), are now on a par with comparable ones of printed/written languages, as exemplified by Kucera and Francis (1967) for adults and Carroll, Davies, and Richman for children (1971), all of which make extensive use of computer compilation and analysis.

Bongers stated that a corpus of at least one million words is required for a valid objective frequency count (Bongers, 1947, page 240). Even with the aid of computers, such a corpus is only laboriously obtained, manipulated, and analyzed. No matter how objective it may be, such a count is always subjective to the extent that someone must select the materials from which samples will be taken, and decide on the size of samples and their method of selection, even if the materials are chosen as a result of consensus of replies to a questionnaire.

A possible alternative to the so-called objective word frequency counts has been suggested by Bernard Shapiro in his doctoral thesis entitled: "The Subjective Scaling of Relative Word Frequency" (1967). Dr. Shapiro determined experimentally that relative word frequencies are a prothetic psychological-additive variable (as are other linguistic items) and that they are best subjectively measured by the "magnitude estimation" technique which follows the Steven's (power) Law. Further studies, if they verify Shapiro's work, may permit the determination of relative word frequencies and the development of relative frequency lists by subjective means and their conversion to objective word lists by means of mathematical formulae, tables, and graphs, thus saving much time, effort, and expense.

The statistical sampling techniques used in the latest Japanese counts also deserve further study in an effort to ensure representative sampling in an economical manner.

SECTION II - DISCUSSION

Purposes of Word Counts

We have seen that word frequency counts have been made in many languages and for many purposes related to teaching and learning; such as stenography, spelling, vocabulary building for graded readers and for determining the essentials of oral vocabulary. They have been made for purposes of psychological research. They have been made on the words used both by children and on those used by adults.

Generally, however, the intent has been to simplify instruction and to economize on time and effort by concentrating on relevant and appropriate materials at successive levels of education whether for the written or oral natural language or some shorthand representation of it.

Active and Passive Vocabularies

It has been determined that there are differences of learning levels to be achieved even within school grades. For speaking or writing an active knowledge is required; in spelling which requires recalling and writing the word in the right combinations of letters, and in talking which requires being able to recall and pronounce the word in an understandable manner with due attention to accent, stress, and intonation. On the other hand, for reading and listening what is required is less complex, namely the visual or aural recognition of the word and determination of its meaning in the context of the other words with which it is found. This is not as simple as it appears, since it also involves recognition of typical sentence patterns of the language involved, but is nevertheless a lesser skill than having to recall and use the words and structures involved as one does in speaking or writing. Active vocabularies, i.e., those used for speaking or writing are referred to by various authorities as "production or expression" vocabularies. Passive vocabularies, i.e., those used for listening or reading are referred to as "recognition, reception, or comprehension", vocabularies.

Items Counted

There are various definitions of the lexical unit to be counted, but in the end, the use of the dictionary word unit appears most efficient, even though in the learning process prefixes, suffixes, inflections, derivatives, and idiomatic expressions must be considered as well as shifts from basic meaning. Vakar defines a word as "every combination of letters with blank spaces on both sides" (1966, Vol. I, page 11).

The first word counts tended to be of the printed or written word, principally the former. The idea was, and still is, that to teach or learn efficiently, vocabulary must be built up in a natural and effective way so that the lexical units which are used most are learned first along with the grammar required to facilitate understanding of the basic structures within which the lexical units appear.

Subjective Discussion in "Objective" Word Counts

The early counts of the printed word were the so-called direct or objective frequency counts resulting in Wordbooks, Word Frequency Books, or Frequency Dictionaries. These counts, although called objective because words were counted and frequencies tabulated, involved numerous subjective decisions which actually made them hybrid objective-subjective counts rather than purely objective ones. Some of the subjective problems which had to be resolved were:

1. What is the purpose of the count?
2. What is to be counted?
3. What is to be recorded?
4. Are homonyms to be counted as one word or as separate words?
5. Are meanings of words other than homonyms, i.e., the semantic subfrequencies, to be considered?
6. How many items are to be counted?
7. How wide a range of categories or material and sources within categories have to be sampled to satisfy the purpose of the count?

8. What should be the time frame of the materials used as sources; i.e., only current materials, or those extending back to a specific date in the past?

9. How will the sampling be done?

10. How many words will be included in each sample?

11. Is frequency to be the only criterion for assigning a rank order of importance to the words determined to be in frequent use?

12. How many and what kinds of items are to be included in the final list culled from among all the items collected?

13. In what formats are the results to be displayed?

Purpose

The first question, the consideration of purpose, sets the stage for answering all the others. However, all of them do not automatically follow from the purpose, since alternative approaches are open. Purposes have been discussed above and many of the more common ones are listed in Appendix 1.

What is to be counted and recorded? This question generally resolves itself into two phases. Initially, everything in the sample is recorded. In the better counts, each word is preserved in context for use in determining variations of meaning. For the purposes of word counting, however, a decision has to be made as to whether to record oral markers, punctuation, exclamations, false starts and repetitions in oral language, obscene words, coined words, ungrammatical utterances, dialectical items proper names, place names, and numbers. The tendency has been to omit markers, punctuation, place, proper names, numbers, false starts and repetitions in oral language and to drop vulgar or obscene words. On the other hand, the tendency has been to convert coined words, ungrammatical utterances, and dialectical items to standard English equivalents. Again, all depends on the purpose. If the purpose is purely to study what is being written or said, then everything can be included, since language is what is being said and written, and not what someone thinks it should be. On the other hand, if the final object is teaching of children, there is little sense in preserving immoral or illiterate expressions for their edification. There has been a long standing tendency to omit place and proper names, numbers, and perhaps days of the week from word lists early in the compilation. The basis has been that general word lists are desired, and that these are either specific--as proper and place names--or so common--as numbers--that they do not belong in the word list. However, at least one modern vocabulary based on a word count, "Fundamental French, 1st Level" includes numbers, days of the week, months of the year, seasons, and measurements as special appendices to the vocabulary on the basis that everyone must use them at some time or other, even if frequency of usage in general conversation or writing is low.

Homonyms and Headwords

There has been a general tendency, until recently, to record homonyms as one word rather than to separate them on the basis of meaning. While this may have been desirable in the earlier counts designed to develop word lists for the teaching of stenography, it it definitely contra-indicated in a study of language, as such. There has also been a tendency to suppress forms and to show in the word lists only the headword with a frequency equal to the total frequencies of all its inflections (declensions and conjugations), derivatives, and unhyphenated compounds. This system has the advantage of keeping the list of basic words short while indicating the frequency with which the basic form of the word appears. However, a better practice is to show the headword and then to indent under it its derivatives and unhyphenated compounds in a word family group, with the total family frequency listed for the headword, and individual frequencies listed for derivatives. A similar problem arises with respect to singulars and plurals. The singular is usually the headword, but often the frequency of the plural is included only in the headword and the plural is never shown in its own right. If we are dealing with words only as simple concepts or ideas, this may have some rationale, but if we are also interested in how the word is used; i.e., whether only or mostly in the singular or plural, subsuming the frequencies under the headword tells the student or teacher only the gross usages of the concept, not the form or forms in which it appears. It would appear best to use the form with the greatest frequency as headword and to indent the plural (or singular) under it indicating its part of the total frequency.

Meaning

A similar, but more important problem arises with respect to meaning. The whole object of language is to convey meaning. To divorce meaning from frequency in lists intended for use in teaching language is senseless. Even Thorndike with the aid of Lorge was finally convinced of this in the 1930's, yet we still find frequency lists coming out with the important element of meaning and semantic frequency omitted. Admittedly, the addition of meaning complicates and lengthens a word list, but it is essentially a part of the word family group of headword, derivatives, unhyphenated compounds, and perhaps inflections. If a word has two or more different meanings which are difficult for the beginner to infer from each other, merely listing the word and its frequency **does not help very much, particularly in speaking** and writing. There should be sublistings indicating the contributing frequency or percentage of total frequency of each of the important meanings of a word. Determining what is important calls for another subjective decision, but, in general, meanings contributing 10 percent or more of the total word frequency should be included. The teacher then has the option of grading his or her materials by teaching initially only meanings amounting to, for example, 75 percent of the total frequency of the word. Without an indication of semantic frequency, the teacher is left to his or her own experience to determine what meanings of the word should be taught and in what order.

Closely related to semantic analysis from a teaching point of view is grammatical analysis. A good count should also show the frequency (or percentage) of total use which each grammatical use of the word contributes since some words function as two or more parts of speech and it is important to a teacher to know whether the word is most used as an adjective, noun or other part of speech.

The problem of whether to list inflections is more difficult. Most languages conjugate their verbs and some decline their nouns and adjectives. Most also compare adjectives and adverbs. A list which classifies all inflections could be very cumbersome, although instructive. At least for a language which does not decline its nouns except for plurals and does not assign a gender to most nouns, the problem is largely one of the advisability of recording and listing the frequency of the conjugations of its verbs. Certainly such voluminous material ought not to be in the lists proper, but putting them in appendices would be appropriate as an aid to the teacher in determining the grading and order of teaching (if worth teaching at all) of the several tenses of verbs, and within tenses the "persons" which are important. It appears that such a listing of verbal conjugations would prove an important economy measure both from the teacher's and the student's point of view. If a verbal form is to be used in writing or speaking only once in a million times, there is little use of teaching or learning it for either active or passive uses except for those who are to become experts in the language; i.e., translators, interpreters or teachers. Without such a list, the chances are that time and minds will be occupied with much excess linguistic baggage to the exclusion of much more important matters.

Quantity to be Counted

On the problem of how many words should be counted, there is empirical evidence ranging from Vakar's Spoken Russian Word Count with a corpus of 10,000 randomly selected running words with 93 sources of one category, (1966) through Eldridges's newspaper count of 1911 with 40,000 running words (Bongers, 1947, page 33) to Thorndike and Lorge's "Teachers' Wordbook of 30,000 Words" (1944) which was based on a combined total of about 23,500,000 running words. Mackey argues that statistically, the greater the number of items counted, the greater the reliability of the counts (1967, page 179). Bongers has repeatedly stated that counts of less than one million running words are of little value (1947, page 240), and Keil (1965) says that the corpus should contain at least ten million running words. Yet many apparently excellent recent counts have far fewer than one million words: e.g., Fundamental French (1st Level) (602,000 total running words with only 312,315 spoken words) (French Ministry of National Education, 1959). Vakar in defending his small 10,000 word corpus stated it was derived from a population of more than one million running words and that "properly conducted random or sequential sampling makes larger word counts wasteful--for after all, the commonest words must be common enough to recur in any text of reasonable length." (1966, Vol. 1, page 10).

On the theory that numbers improve reliability, many researchers--after making their own studies--have added in the results of prior research to supplement or broaden their own. This, in effect, increases both the running words and the number of sources of the composite study. Statistically, this type of additive research should have increased the validity of the final results. However, it must be remembered that the quality of the final study can hardly be greater than the average quality of all inputs in spite of the greater number of words and sources.

Categories and Sources

With respect to range, the question of categories (Mackey uses the term "Registers") of material as well as the selection of sources within categories arises. Special technical counts can be restricted to one category, but a count designed to yield a general vocabulary, especially an active one, must sample a wide range of categories and sources within those categories. Choice of colloquial as opposed to literary style, differences of author style, differences in dialect, and differences of period in which the source was written all affect the occurrence and usage of words. For a good general vocabulary, a wide and current variety of oral as well as printed and written material must be included with a view to deriving words for both active (productive for writing/speaking) and passive (recognition and reception for listening/reading) vocabularies.

The Spanish Vocabulary Count of the University of Puerto Rico (Rodriguez Bou, 1952) also included a category of subjectively selected printed sources based on supplemental texts used in the school systems. This is an indication of the techniques which researchers use in order to ensure that their final vocabulary is representative for the uses to which they expect it will be put: in other words to ensure that the range of their study is adequate for its purpose.

Time Period

The time period in which a source was written will also affect the vocabulary derived from it. Frequency of usage of individual words and their meanings clearly change over time. Here again, the purpose is important in determining range. Josselyn in his word count of printed Russian (1953) could well go back to the mid-1800's since he was interested in a vocabulary to assist readers of literary Russian. However, anyone interested in colloquial oral Russian would use recent sources as Vakar did in taking samples from 200 acts in 93 plays published in 1957 or after. Berger in his studies has found evidence that conversational English may vary considerably over relatively short spans of time and space (1967, page 20).

Source selection was one of the biggest criticisms of Thorndike's "Teachers' Word Books" since in its 1921 edition it leaned so heavily (75 percent of four million words) on the Bible and literary works. As successive editions appeared and additional material was added this situation improved as the biblical/literary sources became diluted in the huge corpus. Nevertheless that part of the original corpus (about 25 percent) which was current in 1921 was already at least 23 years old when the 1944 edition appeared. This combination of the already old and that which became old before the 1944 edition resulted in making at least 15 to 20 percent (four to five million running words) of the final corpus out of date even in 1944.

To overcome the problem of the effects of selection of individual sources on word frequencies, as large a number of categories with as great a variety of sources within categories as possible is desirable within the bounds of manageability and diminishing returns.

Sampling and Sample Size

The number of ways in which samples can be collected from sources is almost infinite. To reflect the structure of the language and word meaning, however, if the material is taken from printed or written material, it should seldom be less than sentence length and paragraph length might even be better, regardless of whether the material is taken sequentially, randomly, or according to some other predetermined pattern until the required number of running words has been taken from each source.

Sampling of oral language can be done in several ways. In any method, spontaneity is desired. Cues in the form of words or subject areas are used to evoke the flow of words. The number of words taken from each informant depends on the size of the corpus desired and the number of informants. It is important that categories of subjects to be talked about and stimulus words or pictures used to elicit responses of either connected discourse or individual words be selected in advance in order to ensure thorough coverage of the several aspects of daily life. Since each person speaks at his own rate it is difficult to determine how long the informant has to talk to provide his quota of words. However, the common research practice has been to record discourse by tape recorder for periods running from 3 to 12 minutes.

If discourse on a subject area is desired, the informant is asked to talk, for example, about his job, his family, his home, the furniture in his home, his hobbies, or sports, as desired. If the speaker slows down because he is running out of subject matter or ideas, he may be prompted by asking leading questions so that he will touch on aspects of the subject he has overlooked.

Another method of evoking spontaneous speech has been to use the Murray Thematic Apperception Test of 1943 or similar device. Informants are asked to talk about the pictures which are the basis of the test, being careful to match the picture with the sex of the informant in order to obtain subject related words from individuals most likely to be well acquainted with the subject.

Another method is that of association; either free or controlled. In free association, the informant is asked to write down everything that comes into his mind during a specified time period, e.g., five minutes. In controlled association, the informant is given a stimulus word and asked to write down words, e.g., all the nouns, verbs, and adjectives, which that stimulus suggests to him in a specified period of time. Free association has been used at least since 1936 when Buckingham and Dolch published "A Combined Word List".

Still another method, useful with school age children for written counts, is to examine written compositions at various grade levels on various subjects.

Variations and combinations of these techniques have been used widely in the past 20 years in English, Spanish, French, and German word counts, and have yielded, particularly for oral word counts, as useful results as extensive sampling has for the written or printed counts of the past.

The stimulus words, related to subject area, for example, have been quite useful in discovering the so-called concrete, topical, or utility words. These are commonly nouns or adjectives. They are often single meaning words which relate only to specific things and are not likely to occur in general frequency counts unless the count has a very large corpus which happens to contain a contribution from a source related to the subject in which the concrete word is likely to appear. Many of these words in a general count would have a frequency of only one in several million and a very restricted range, yet they are absolutely essential if one wants to talk about the specific subject to which they are related. Many of the frequency counts were meant to produce vocabulary for teaching purposes, but failed to produce words such as chalk or chalk (black) board which are intimately related to the classroom. Until recently, the only way to obtain such essential words was to compile a very large corpus or to use subject or interest area convesational topics, controlled associative techniques or both. Recently, however, Richards (1970) has recommended that a system called "Word Familiarity" which is a subjective rating of a list of words according to the relative frequency with which informants believe they expect to encounter the words.

Relative Importance of Words

Once the different words (types) have been selected from the running words (tokens), and the frequency count completed, the question arises as to the relative importance of the words which have been isolated. On this depends the order of consideration and presentation if one is preparing a textbook for teaching. Originally, the raw frequency was taken as the indication of the relative importance of the word; the higher the frequency the more important the word. This appeared to be true for teachers and students of stenography and spelling for whom many of the early counts were made. However, the criterion of frequency by itself became suspect when studies were made of the frequency of the words in each of the sources which contributed to the total frequency. It was discovered that some words were fairly evenly distributed among sources and thus truly general and useful regardless of subject matter. Such words included, but were not restricted to, the so-called structurals or functors. A problem in relative word importance became apparent, however, when it was discovered that a word with a high total frequency might be derived from a single source or small number of sources out of all the sources which contributed to the total corpus. **Unevenness** of frequency distribution across sources is generally an indication that the word is somehow specific to one or a few subjects and is not encountered generally. (Bongers has labeled such types "environmental" words.) The problem of relative importance or rank order of words on vocabulary lists was thus broadened to the question of what comes first; total frequency or range (number of sources in which the word appears with consideration of the frequency in each source). Some early researchers opted for frequency, some for range, and many others adopted various objective and subjective formulas for combining the two in determining word importance. Most continued to use frequency as basic but used a method such as dropping words that occurred only in one or two sources as not being representative enough for consideration, at least in basic or beginning vocabularies. Bongers has held that trying to correlate frequency and range is an impossible task, since after about 1200 words which are common to most subjects, the words obtained in any word count are so highly dependent on selection of sources that no meaningful permanent relationship between frequency and range can be derived.

Bongers concluded that the best method to date of dealing with the frequency/range problem is that developed by de la Court, while working in Indonesia. De la Court's system was briefly this: His total corpus was one million. When he had counted 500,000 words, he totalled the frequencies of each individual word. He similarly totalled the frequencies of each individual word in the second half of the corpus. Then he compared the frequencies for each word in each half of the count. If they were far out of balance, e.g., had a ratio of 1/10 or less, he dropped the word as being too specific for a general list. He dropped 26 words for this reason alone. Vander Beke in his French Word List dropped any word not appearing in at least half of his sources, thus eliminating many concrete nouns.

Ernest Horn in his Basic Writing Vocabulary (1926) argued that there are two measures of importance in judging subject matter for inclusion in lists: frequency, and value attached to each occasion when material is needed or used. The value attached to each occasion according to Horn was an indication of two types of range; geographical and across writing samples. In effect, Horn obtained range estimates based on number of types of correspondence in which a word was found, the numbers of writers who used it, and also where the writers lived. Bongers did not disagree on distribution as it relates to use in the sources employed by Horn but he believed that Horn's worry about geographical distribution was a sampling problem which could have been handled by judicious selection of sources and increase in their number.

Faucett and Maki (1932) in "A Study of English Word Values Statistically Determined from the Latest Extensive Word Counts," tried a different system in an effort to combine the range and frequency ratings in the Thorndike and Horn Lists. They placed all the words on a scale from 1--the most valuable words to 120--the least valuable, on the basis that words of value 1 had the widest range and greatest frequency and 120 the least of both. Intermediate groupings were defined as: 1 to 9: Indispensible; 10 to 34: Essential; 35 to 80: Useful; and 81 to 120: Special. One problem that they encountered arose from the purpose of the Horn list. Horn dropped all words spelled with three letters or less, thus suppressing a large group of short words. This obvious attempt to deal with the difficult problem of resolving frequency and range failed because Thorndike and Horn didn't use the same type of frequency ratings and did not even agree on their definitions of a word.

More recent studies indicate that relative word importance, at least for language teaching purposes, depends on factors other than frequency and range. These factors include availability, coverage, and learnability. We have already discussed "availability," otherwise known as "utility", when we discussed the concrete, subject-oriented nouns and adjectives. However, verbs too can be situational or specific. For example; lists made of the ten most frequent and the ten most available French verbs revealed only one word common to both lists, and that was "aller" (Mackey, 1967), which may be glossed as "to go" but also "to be going to" or "about to do" something. It was fifth on the most frequent list and seventh on the most available list. The only reason that it was on both lists was that as a "content" verb--to go--it was used frequently, and as a structural (auxilliary) verb it is popular as a present progressive auxiliary indicating an action to be taken in the near future. In English, "go", "got", and "have" have the same duality of function.

With respect to "coverage", the importance of a word depends on its ability to replace the greatest number of others with which it is wholly or partially synonomous. In a basic learning vocabulary, it is preferable to include one word that can be used to serve for six others. In this way, the learning effort is reduced to about 10 to 15 percent of what it would be if all seven words had to be learned. Coverage can be broken down into four subdivisions: Inclusion - a general word which includes the meaning of several more specific ones is to be preferred to a word with only one or two meanings; Extension - a word with many full or partial synonyms is preferred to one with few; Combination - a simple word which can be used to combine with others in compounds which replace other individual words is preferred to a word which does not combine often in general usage; and Definition - words that are most useful in defining--and, therefore, substituting for others is to be preferred to one which is of little use in explaining other words. Michael West explores this property of words in his Definition Vocabulary (1955) and Ogden in the development of his Basic English (Graham, 1968).

J. G. Savard in his book entitled La Valence Lexicale (1970) discusses the use of word coverage as an alternative to frequency in determining the relative value of words. He found that the correlation between word coverage and frequency is weak. He believed that they are two very different principles related to word value, and that word coverage was no less valid a measure than frequency. He recommended that studies be continued to try to find correlations between frequency, range (distribution), availability (utility), and word coverage with its four constituents as listed above. He believed that word coverage represents a new variable which should be considered in determining the assistance a select vocabulary can render to language learning. He argues that it, or something like it, is needed to supplement frequency, range, and availability in determining the relative value of words constituting limited vocabularies. Word coverage, in his opinion, is a measure of verbal economy and will have useful side applications in the development of dictionaries, glossaries, and thesauruses.

A word is important from the point of view of "learnability", which is an awkward way of saying it is easier to learn than other words. Logically, "learnability" may be considered to be a function of "similarity", "clarity", "brevity", "regularity", and "learning load" or "burden". Similarity generally occurs because words are cognates in the two languages concerned; they are generally orthographically and referentially similar. This is not always true, however, because words may be more inclusive in one language than another and their frequency of usage will likewise be different. *Clarity* is usually found in concrete as opposed to abstract words. *Brevity* is a function of being spelled with a few letters and being easily pronounced. *Regularity* is a function of following grammatical rules, such as regular plurals or regular conjugations. Regular words are preferable to irregular ones which require more learning effort. *Learning load* is inherent in the preceding four aspects of learnability; words selected on the four preceding criteria will normally be easier to learn. Unfortunately, words which are learned easily may not be the most useful.

Swenson and West in their study "On Counting of New Words" (1934), included "A Set of Rating Scales" which are comprehensive in their enumeration of the gradations of difficulty of learning idioms, cognates, compounds, spelling variations and semantic shifts of words. Because of their special interest, these scales are reproduced in full in the following displays.

Ernest Horn in his *Basic Writing Vocabulary* (1926) also considered spelling difficulty as an input to word importance, but his purpose was quite different. Although the writers of basic vocabularies are looking for balanced word lists they want their words to be as simple as possible and still permit expression at an adequate level of language proficiency. This motivation leads them to delete a word that is in some way difficult if a simpler substitute is available. Horn, on the other hand, in trying to help teachers with their job of getting pupils to learn to spell well, used the greater degree of difficulty as a rationale for including a word in his list.

BEST COPY AVAILABLE

Scale I. A Rating Scale for Changes of Meaning

? Perceptible ?		0	A word already learned, and now used in exactly the same meaning.
	1		
		2	A change of meaning just perceptible to the teacher; query worth pointing out?
	3		
		4	A change of meaning so slight that it would not be noticed in reading—but it might be pointed out in speech.
	5		
? Inferrable ?	5		
		6	A change of meaning which would be noticed in reading and would cause a moment's hesitation.
	7		
		8	A change of meaning which would be noticed in reading and may cause considerable hesitation, but the meaning will be inferred by all the pupils eventually.
	9		
		10	The average child might *just*—or just not—be able to guess the meaning in reading. **MID POINT**
	11		
? Explainable	11		
		12	The new meaning would probably not be guessed in reading but is easily grasped when explained, and is easy to explain.
	13		
		14	The new meaning could not possibly be guessed in reading but can be explained—with medium difficulty.
	15		
		16	It is difficult to show the connection between the old and the new meaning; the connection is barely perceptible.
	17		
? Twistable ?	17		
		18	Almost a homonym. The old meaning can barely be twisted into the new one.
	19		
		20	An entirely new word of average difficulty which can readily be translated into the mother-tongue by one (or two) equivalent words.
	21		New words which are less readily translatable.
		— 30	An untranslatable word; it needs a lecture to explain it.

BEST COPY AVAILABLE

EXAMPLES OF RATINGS ON SCALE I.

1. The discovery of America: the discovery of X Rays. The middle of the room: the middle of the night.
2. A loud cry: a loud voice. An old man: an old building.
3. One minute to six: wait a minute. High up the hill: high up in his profession.
4. A bad boy: a bad egg. Hollow (adj): in the hollow of a tree.
5. The sun is in the sky: standing in the sun. Trees growing about the house: there were no people about in the streets.
6. A debt of $1000: I am in debt to him for his help. To fight against the enemy: I am against any change in the law.
7. Peter loves Jane: I love sausages. He has a weak heart: a kind heart.
8. Of equal size: my equals and my betters. Neck of a man: neck of a bottle.
9. To touch with fingers: leaves touched with gold. The eye: to eye.
10. Form (=shape): form of proceedings.
11. Hollow: near the wood there is a beautiful hollow. Match (marriage): to match colours.
12. Hand: he writes a good hand.
13. After hearing what he said: her hearing is bad. They had an argument: that is a strong argument for.
14. To touch: touching.
15. A room: room. Air: (-manner)
16. Arch: archer.
17. If he does I shall . . .: go and see if she is ready.
18. A meal: meal.
19. Arm (part of body): arms.
20. A match (light): match (marriage).

BEST COPY AVAILABLE

Scale II. A Rating Scale for Idioms

? Perceptible ?	1	0	The idiom offers no trace of difficulty (and is exactly the same as that of the mother-tongue). Would not be noticed as an idiom.	
	3	2	A very obvious and self explanatory idiom (almost the same as that of the mother-tongue). Might not be noticed as an idiom.	
	5	4	The meaning is very clear, but not quite obvious. Would probably be noticed as an idiom.	
? Inferrable ?	5 7	6	Would be noticed as an idiom, and would cause a moment's hesitation.	
	9	8	Would cause considerable hesitation, but the meaning would eventually be guessed by all.	
	11	10	The meaning of the idiom might, or might not, be guessed by the average pupil.	**MID POINT**
? Explainable ?	11 13	12	The meaning of the idiom would probably not be guessed by the average pupil, but is very easily explained.	
	15	14	The meaning could not possibly be guessed by any pupil, but can be explained with medium difficulty.	
	17	16	The meaning is just perceptible in the words when they are explained,—but is difficult to explain.	
? Twistable ?	17 19	18	The words can just barely be twisted into the meaning.	
		20	The idiom is quite unexplainable: the whole idiom has to be taught as one word.	
	21 — 30		Idioms in which the pupil is especially liable to go wrong, e.g. such as mean something different if translated literally into the mother-tongue.	

BEST COPY AVAILABLE

EXAMPLES OF RATINGS ON SCALE II

3. Put to death before the eyes of his friends. War came to an end.
4. That's new to me. Nothing on earth would . . .
5. A button has come off my coat. The building was in flames.
6. Much in request as a singer. Work on hand.
7. Outside the field of his interest. Give me a hand with this box.
8. To go there on foot.
9. Between now and then. Keep to the rules.
10. Can have anything in reason. To come into line.
11. I can't put my hand on the paper I want. Worked by electricity.
12. Go wrong. I sent for the doctor.
13. He turned up his trousers. Six years old.
14. He came in person. A run on the bank.
15. Go on singing. I look forward to the party.
16. Far nicer. To make money (by selling books).
17. To put up a good fight. Set out.
18. In good order. In order to. He let me down.
19. Said it with his tongue in his cheek.
20. So long! Egged on.

BEST COPY AVAILABLE

Scale III. A Rating Scale for Cognates

(Weight for spelling-pronunciation is to be added to cognates, where necessary. Words cannot be rated as cognates unless the native word is known to the pupil).

? Obvious ?

0 Such perfect identity of form and meaning that the word is not noticed by the pupil as being new.
1

2 There is a just perceptible difference of form and/or meaning —but the sense is very obvious.
3

4 A fairly obvious relationship; the change of form and/or meaning will be readily understood.
5

? Inferrable ?

5
6 Such difference of form and/or meaning as will cause a moment's hesitation,
7

8 Such difference of form and/or meaning as may cause considerable hesitation, but the cognate will eventually be identified and interpreted by all.
9

10 A cognate which just might—or might not—be identified and interpreted by the average pupil. **MID POINT**
11

? Explainable ?

11
12 The cognate would probably not be identified or interpreted by the average pupil, but will be very readily grasped when pointed out.
13

14 The cognate could not possibly be identified or interpreted by any pupil, but the relationship of the foreign to the native word can be explained with medium difficulty.
15

16 The relationship of the foreign and native word is only just perceptible and is very difficult to explain, but the cognate is probably helpful.
17

BEST COPY AVAILABLE

EXAMPLES OF RATINGS ON SCALE III

(French words—English)

0. Courage—Courage
1. Émotion—Emotion
2. Un toast—Toast
3. Flanc (d'une montagne)—Flank
4. A bord—Aboard
5. Compter—To count
6. Parent—Parent
7. Bravoure—Bravery
8. Se moquer de—Mock at
9. Cave—Cave
10. Vous avez raison—Reason-able.
11. Partie—Party
12. Se dresser—Right dress!
13. Parcelle—Parcel
14. Rude—Rude
15. Trouble—Trouble
16. Spirituel—Spiritual
17. Pavilion—Pavilion
18. Figure—Figure
19. Gallanterie—Gallantry
20. Adresse—Address (on a letter)

BEST COPY AVAILABLE

Scale IV. A Rating Scale for Compounds of Known Elements

N.B- The Prefixes and Suffixes are scored as new words on their first occurrence; this scale refers to subsequent compounds only.

? Perceptible ?

0 Addition of an absolutely invariable P or S which never causes any change of form or meaning.
1

2 Addition of a regular P or S with such change of form or meaning as would hardly be noticed in reading.
3

4 Addition of a P or S with such slight change of form or meaning as will probably be noticed—but it will cause no difficulty in reading (but must be pointed out for speech).
5

? Inferrable ?

5
6 Addition of a P or S with such change of form or meaning as will cause a moment's hesitation in reading.
7

8 The change of form or meaning may cause considerable hesitation, but the word will be identified and interpreted by all the pupils eventually.
9

10 The average pupil may just—or just not—be able to identify and interpret the compound in reading. **MID POINT**
11

? Explainable ?

11
12 The compound would not be identified or interpreted in reading, but is easily grasped when explained, and is easy to explain.
13

14 The compound could not possibly be guessed in reading but can be explained,—not easily, but without great difficulty.
15

16 The meaning of the P or S and original root are only just perceptible in the compound, and the compound is not easy to explain.
17

BEST COPY AVAILABLE

? Twistable ?

17

18 The meaning of the P or S and root can just barely be twisted into the meaning of the compound. It is doubtful whether analysis is useful.

19

20 The P or S creates a new meaning bearing no relation either to the original meaning or to the P or S.—Or the P or S changes its meaning so widely as to amount to a new P or S (rated as a new word).

21 — 30 Compounds which cause special difficulty, e.g. such as tend to be misused in the literal sense though the real meaning is far different.

EXAMPLES OF RATINGS ON SCALE IV

1. Childhood. Button-hole†
2. Leader. Dangerous
3. A prefix. Tidiness
4. Imprison. Boundless
5. Transplant
6. Nonsense
7. Terrify. Misadventure
8. Wasteful
9. Irrecoverable
10. Money-order

†(not 'flowers').

11. Enlist
12. Underling
13. A falsehood
14. A two-seater Moreover
15. Transact
16. Profiteer
17. Mislay
18. Homely
19. Anchorage. Engender
20. Well-off

BEST COPY AVAILABLE

Scale V. A Rating Scale for Spelling-Pronunciation Discrepancy

This scale is not intended to assess the relative difficulty of the fundamental sounds of the language, but merely such tendency to mispronounce or misspell as is induced by a spelling which does not correspond to the actual sound of the word.

(Half this rating is to be added as extra weight to the rating of any new word, viz. any group of letters or sounds not previously encountered. No word is to be rated twice for spelling-pronunciation, even if the first appearance of the word was in a totally different meaning.)

(1) The word is pronounced just as it is spelled, and spelled just as it is pronounced; no possibility of error.

(2) A slight divergence which may lead to error in pronunciation or in spelling.

(3) A less easy or safe word, but still below the average of those that give any trouble.

(4) If the word were dictated to an average class, without previous experience of it, nearly half the pupils might misspell it; or, written on the blackboard, nearly half the pupils might misread it.

(5) If the word were dictated to an average class, more than half the pupils would make a mistake.

(6) A definitely troublesome word, but not among the worst.

(7) The notorious trouble-givers.

(8) One of the most often quoted absurdities of English spelling; a word that almost all foreigners misspell—or, if they spell it right, they mispronounce it following the spelling.—Also words in which the pupil tends to be misled gravely by the spelling of a word in his mother-tongue.

Examples

1. This. Time.
2. Blade. Dress.
3. Shock. Roll.
4. Separate. Soup.
5. Science. Soap.
6. Beautiful. Doubt.
7. Scythe. Touch.
8. Cough, tough.

The selection of words and format in a vocabulary list is not too difficult if all the five criteria (frequency, range, availability, coverage, and learnability) are in agreement. The problem arises when the criteria are in conflict on the selection of a word. Any one of the criteria could be in conflict with all the others just as easily as it could be in agreement with them. This means that in case of complete conflict some 10 conflict areas have to be resolved. The final resolution depends largely on the uses to which the list is to be put. For a combined use, such as both a speaking and reading vocabulary, Fries and Traver (1950) suggest order of precedence among criteria as follows: frequency, coverage, range, availability, and learnability.

The foregoing yardsticks for measuring importance or value by no means exhaust the list. As intimated above, almost every word counter has had his own system, either original or a modification of an earlier system used by another word counter. Perhaps the best list of considerations other than those discussed above is to be found in the "Interim Report on Vocabulary Selection" (Carnegie Report) of 1936 (Michael West et al.). It listed as possible criteria: frequency, structural value (functional types), universality over wide geographic area (like Horn); applicability to a wide variety of subject matter (general use words); value for purposes of defining other words (West's Definition Vocabulary); value for word building (ability to combine into compounds, discussed above); and stylistic function (use to express precise meanings and in conversation).

Lawfulness of Vocabulary Distributions

There must be a balance between comprehensiveness and usefulness. Thorndike, with the aid of Lorge in the later stages, worked his original list of 10,000 words up to 20,000 and then to 30,000 words over a period of 23 years. The Thorndike Word Book was certainly comprehensive but in reality not as useful as it appeared because of its antiquated sources. Thorndike counted and borrowed over 23,000,000 words for his 30,000 word list. This tremendous corpus size and others like it have led Frumkina (1964) to propose an application of Zipf's Law which would allow calculation of the corpus size required to provide a word list which will be statistically valid down to a pre-selected frequency within a predetermined margin of error. Using the large corpora assembled by other investigators, Frumkina attempted to estimate the population values for the frequencies of individual word types within a specified interval of the sample spaces, these estimates being based on the observed regularity of the frequency-rank relationship discovered by Zipf. More recently, Carroll (1971) has used a log-normal function to estimate the population values of the word types from the data of the American Heritage Wordbook.

In contrast to the comprehensive word lists of the Thorndike variety are the so-called Basic Vocabularies for the teaching of foreign languages. These have tended to run from 600 to 3000 words for a good four-year high school language course. However, these numbers may be deceptive, depending on how "word" is defined. If the dictionary entry is taken as the "word", but is deemed to include within it all its inflectional forms, derivatives, compounds, and the semantic variations of each, a list labelled as 1000 "words" may actually be effectively a 6000 word list as far as the learning effort required of the student is concerned. The point is that the short basic vocabulary, although built up of word families with related forms and meanings, may give a false impression of the effort required to learn the fundamentals of a foreign language. Given a reasonable balance between subsuming all forms under the headword and separate entries for each form, most researchers agree that 3000 to 6000 words provide a good basic vocabulary.

The location of the point of diminishing returns as applied to frequency has been an object of controversy for several years. Ernest Horn (1926) in his Basic Writing Vocabulary said, with respect to spelling, that after the first 1000 words, the addition of each group of 1000 words in a spelling list adds a very small percentage to the number of running words that one can spell. For example, the person who knows how to spell the 4000 commonest words can add only a little more than one percent to the number of running words he can spell by learning an additional 1000 words, since the new words are those of low frequency of occurrence. West suggests that a general vocabulary of 7000 words will enable a person to read most novels, and that for speaking, an individual needs a vocabulary of about 2800 general words. After reaching the 7000 and 2800 word limits, the person must start learning specialized vocabularies in his field(s) of interest. These figures indicate that the 3000 to 5000 word vocabularies are valuable, even though they may have to be graded into smaller increments for instructional purposes.

Another problem is how to determine the words to be included in a useful reading or speaking vocabulary of the size cited by West, above. Most authorities, including Ayres, Thorndike, and Horn have concluded that objective word counts must be extremely extensive in numbers of running words and must sample a very wide range of categories and sources to be accurate beyond the first 500 to 1500 words. Ogden and Palmer believed that word lists of equivalent length could be compiled subjectively with the same accuracy. In objective counts, after the first 1500 words, the lists tend to reflect the subjectively chosen sources and categories, whether the count is made of oral, printed, or written language. Nevertheless, Thorndike held that his 1921 list of 10,000 words was good enough for educational purposes through the first 5000 and that it was generally useful throughout its full 10,000 words.

Although direct comparisons are impossible because of the different methods used and definitions applied, most researchers have agreed that only a very small number of high frequency words are actually used in the majority of writing and even fewer in speaking. Jones and Wepman (1966) found that 33 spoken words used by adults accounted for more than 50 percent of all the words they recorded. Herdan reported that for printed English, the 67 most common words accounted for 50 percent of all words counted. A study of the Thorndike-Lorge 30,000 word list by Jones and

Wepman concluded that 89 words accounted for 50 percent of all words used in printed English. (The differences between Herdan's conclusion and that of Jones and Wepman may be caused by differing approach, but it might mean that stylistically, printed English is becoming more laconic, since Thorndike's material on the whole is rather old.) At higher percentages, Eldridge (1911) found that 750 words constituted 75 percent of words generally used in newspaper English. Cook and O'Shea (1914) found that 763 words constituted 90 percent of words used in correspondence (but 42 percent of the 763 were highly repetitive function words). Dewey (1923) concluded that 1000 words constituted 75 percent of words generally used in printed American English. The Bell Telephone System (1930) calculated that 700 words constituted 95 percent of all telephone conversations. D. B. Johnson (1972) reported that the most frequent 2000 words in Czech, English and Russian account for between 75 and 80 percent of words normally used in print and that 5500 to 6000 words will include over 90 percent of general reading material. Johnson's studies, thus confirm, in general, Horn's remarks on the point of diminishing returns near the 4000 mark and West's opinion that 7000 words are required for reading novels comfortably.

The foregoing figures would indicate that a 1000 word vocabulary is a good starting point, but a more representative list like the Thorndike 10,000 word list is necessary to provide the less frequently used words required to bring the student up to proficiency in reading and speaking, as defined in terms of vocabulary size by West and Johnson. Palmer believed that it is best to have a limited general vocabulary as a base and to supplement it with lists designed for specific technical, vocational, and academic fields. Experience with Fundamental French (1959) and Basic (Spoken) German (1964) appears to support that belief.

Disposition of High Frequency Function Words

A problem that has to be resolved with respect to word lists derived from frequency counts is what to do with the most frequently occurring words which invariably appear in the first 500 words of any frequency count of a given language. These are the so-called grammatical (structural or relating) words with which we speak or write and for this reason appear to be largely **independent** of subject matter. The content words, i.e., what we talk about, are largely nouns and verbs, many of which have very specialized meanings. They are, therefore, highly situational and, in general language, have a low frequency of occurrence. As a result of the high frequency and constant appearance of the structural words, word frequency counters have as a matter of routine deleted from 50-300 of the most common ones from their frequency counts and have placed them in separate lists or appendices. These are words which experience has shown will appear with the highest frequency and, therefore, early in any frequency-based vocabulary count of the language. They are listed separately in order that the main lists may concentrate more on the "content" words of the language. Also separately listed by most counters, but for reason of their specialized use, are cardinal and ordinal numbers as well as proper names and place names which are so subject or area related that they warrant no place on a general word list.

Word Groups or Collocations

Another problem which has to be addressed in frequency counts is that of what are commonly called "idiomatic expressions". Some special meanings appear to depend on the combined sense of a more or less fixed association of words which have come to convey a meaning separate and distinct from the sum of the meanings of their component words. Palmer made a study of "idiomatic" expressions and arrived at the conclusion that the term was actually a misnomer. In the "IRET Second Interim Report on English Collocations" Palmer reported on his study of the overlapping fields of vocabulary and syntax. The so-called "idioms" fall into linguistic groupings which Palmer called "Pliologs" or "something more than words". Within Pliologs he distinguished among "linguistic formulas" (conversational expressions, proverbs, aphorisms, and quotations), "syntax patterns" (mainly grammatical), and "collocations" proper, such as verb-, noun-, adverb-, and preposition-collocations. However defined, these word groups are an essential aspect of language proficiency.

In the 1929-30 period, before Palmer published his study, three "idiom" lists were published under the auspices of the Canadian and American Committee on Modern Lanugages or the American Council on Education. The first was Hauck's on German, which supported B. Q. Morgan's "German Frequency Book" with 959 idiomatic expressions based on a minimum frequency of two and a range of one. The second was Keniston's on Spanish with 233 entries which were checked against Buchanan's "Graded Spanish Word

Book". It placed primary emphasis on range as a criterion for idiom selection, using three out of a hundred as the cut-off point. The third was Cheydleur's on French with 1724 entries with a minimum range of three out of 87 sources.

Sometimes idiom lists such as Hauck's and Keniston's above, were not only checked against or designed to support a specific word list, but were derived at the same time. An instance of the latter is de la Court's collocations which were a part of his list called "The Most Frequent Dutch Words and Collocations". It had a list of 3296 words and about 2000 collocations as defined by Palmer, above, in his second IRET Report. The so-called "Linguistic Formulas" (sometimes called Category II items) such as proverbs were not included since none attained the cut-off frequency of five.

In German, two more idiom lists are important; Purin's and Pfeffer's, with Pfeffer's being by far the more important. Purin's "A Standard German Vocabulary of 2932 Words and 1500 Idioms" is a secondary list derived from prior vocabularies and frequency counts (1937). It is of interest since, like de la Court, Purin recognized that basic concepts and meanings are often conveyed by idiomatic type expressions and deserve recognition as a part of vocabulary. It is also of interest since in it the meaning of the idiomatic entry was frequently illustrated by using the idiom in a contextual utterance.

Perhaps the best of the current idiom lists is Dr. Pfeffer's "A Spoken German Idiom List" (1968). It is third in his excellent series of Spoken German. Pfeffer does not refer to Palmer's studies on idioms, but he does refer to the Hauck, Keniston, and Cheydleur lists mentioned above. He defines his idioms as "semantic restrictions of syntactically collocated parts" in which varying degrees of restriction may occur. The Pfeffer list was derived with the aid of computers from the research done to produce his "Basic (Spoken) German Word List" (1964) and "English Equivalents" (1965). Pfeffer selected 1026 idioms from the oral material of his Word List with a frequency to range ratio of 3/2 or higher and 99 others which were discovered while developing his topical (utility or available) words and while rounding out his Word List by empirical additions. All of the words composing the 1026 idioms are found in his "Basic Word List". The total of 1125 idioms represent about 85 percent of the restricted forms and related patterns (idioms) found in spoken German.

Need for Uniformity

It is apparent in reviewing the history of word frequency counts and related vocabularies or word lists that the methods and techniques are as varied as the researchers and their purposes. Now that the science of word counting is evolving rapidly, with the oral count coming of age as electro-mechanical techniques of recording speech have become available and both oral and written/printed counts becoming subject to manipulation by computer, it would appear that we need a new convocation of word counters similar to those sponsored by the Carnegie Foundation in 1934-35. The purpose would be to coordinate the efforts of the many researchers by exchange of information, deciding on definitions, and discussion of the relative merits of the several methods and techniques being used to arrive at the many subjective decisions which have to be made. It is precisely these differing techniques, methods, and subjective decisions that make much of the research so diverse as to make comparisons impossible without considerable manipulation.

Evidence that this need for uniformity is, and has been, felt is found not only in the Carnegie Conferences under the leadership of West, but also by the observations of others who have been frustrated in their attempts to grasp the status of the developments in linguistics and language teaching because of the lack of comparability of the efforts of previous and contemporary researchers in the field. Such lack of uniformity has occasioned extensive efforts to recast the work on one study in terms which will make it comparable with that of another. These are required to avoid invalid comparisons or simple inability to find common ground. Rolf-Dietrich Keil of Germany has recently addressed this subject and made a passionate plea for standarization in his "Einheitliche Methoden in der Lexikometrie" (1965).

Formats for Display of Results

The basic format of most frequency counts has been to list the selected number of words or other items in order of frequency and then list them in alphabetical order. An expansion, as range or distribution began to be considered, was to list the selected words according to frequency and to add the range in a parallel column, or vice versa depending on whether frequency or range was considered the more important. An alternative relative ranking can be obtained by any of the various formulae combining frequency, range and other value judgements into some numerical index representing word importance. This composite value is used to determine the order of listing of words. Total frequency and range are then listed in parallel columns for each word. A refinement of the above is to add columns for each word indicating its range and frequency in each of the categories of material from which the count was compiled. A final refinement is to add a frequency count of the grammatical uses of each word, i.e., how many times it was used as a noun, verb, adjective, or adverb.

In the final analysis, the format for display of results depends on the purpose of the count and the uses to which it is expected it will be put. Usefulness to the reader is the most important criterion. With the assistance of computers, the variety of formats of display of material has increased enormously and there is little reason, from the point of view of time, not to present the material in its most useful form for one or several groups of consumers.

Value of Objective Word Counts

There are still some who, like Palmer in the 1930's, believe that objective counts are useless or, at best misleading, or if they do produce anything it is only a passive (reception) reading or listening vocabulary. One of the more critical articles is W. E. Bull's "Natural Frequency and Word Counts" (1949). The subtitle "The Fallacy of Frequencies" is a good indication of the tenor of his article. Bull argues that:

1. There is an inverse relationship between natural frequency of a grammatical form (such as a noun, verb, article, or adjective) and the frequency with which each form is used. This is borne out by the high recorded frequencies of the relatively few grammatical (functional or structural) words (articles, adjectives, pronouns, prepositions, conjunctions and relating verbs) which provide structure to the language, and often have multi-meanings, although they are not content-bearing words. On the other hand, the real "content" words which convey the meaning of what we talk about tend to have fewer meanings per word, perhaps only one, and refer to specific objects and situations.

2. Any word count is statistically valid only for what is included within it. Keil recommends at least ten million words (1965). Variation in corpus selection does make a difference in the words discovered. That fact is reflected in the decreased comparability among frequency lists after the first 1000-1500 words.

3. Extremely high frequency words are rarely the content-bearing elements of any communication.

4. Range and frequency are determined by two different forces; linguistic and cultural.

5. It cannot be assumed that there is a correlation between frequency and utility. This depends on what is meant by "utility"; a structural word is being used grammatically and necessarily so. It, therefore, has "functional utility", but it may not be used to convey the real cultural meaning of the utterance and, therefore, lacks "concept conveying utility". That this observation is true is substantiated by the need to discover utility (available or topical) words by "centers of interest", "topical subjects" or other methods used in developing "Fundamental French (1st Level)" (French Ministry of National Education-1959), the "Basic (Spoken) German Word List (1st Level)" (Pfeffer-1964), and the "Puerto Rican Spanish Vocabulary Count" (Rodriguez Bou-1952).

6. There are so many factors and uncontrollable elements in life and language that no satisfactory results can be obtained by attempting to reduce such natural heterogeneity by statistical methods. Word counts cannot be considered a valid representation of a people's culture and linguistic activities. As a result, their pedogogical usefulness is extremely dubious.

In general, Bull's arguments at least point up the kinds of questions which need to be addressed if frequency counts are to be useful. However, Bull's overall condemnation of word counting is too strong when one considers the better modern (modified objective) research methods such as those employed by Pfeffer since the early 1960's in his continuing study of

German. Dr. Pfeffer is using for spoken German, a combination of nearly spontaneous speech on general subject areas, topical areas of interest to elicit available and utility words on specific subjects, and empirical (pragmatic) examination and comparison of the results of the first two methods to supplement his word lists. Aided by computers, he has proceeded from his basic word list to semantic classifications with their English equivalents, and has, thereafter, isolated semantically restricted combinations of words in his "Idiom List" (1968).

Dr. Pfeffer's improvement on the Palmer formula of objective, subjective, and pragmatic procedures for developing vocabularies is encouraging. Coupled with sophisticated measures of word importance as developed by Mackey (1967) and his associates at Laval University in Quebec, the Pfeffer research should result in extensive and profitable pedogogical use in the teaching of German and, by transfer, in the teaching of other languages, in spite of Bull's earlier pessimism (1949).

Pfeffer's study of Spoken German, together with "Spoken Russian" (Vakar-1966 and 1969) and "Fundamental French" (1959), coupled with the Wepman and Hass children's count (1969), the Jones and Wepman adult count (1966), the Howe's adult count (1966), the Beier, Starkweather, and Miller children's count (1967), the Berger count of conversation (1967), and the Black and Ausherman college student speech count (1955), have given impetus to studies of the spoken language. Comparative analyses of the difference between conversational oral speech (Berger) and more formal classroom presentations (Black and Ausherman) can and should be made. Concurrently, however, we should also maintain the impetus of work in the field of printed and written language as illustrated by Kučera and Francis (adults-1965) and by Carroll, Davies, and Richman (children-1971).

Areas for Further Research

At the same time, there is a need to explore new fields, such as those indicated by Richards and Shapiro. Richards (1970) developed the concept of "Word Familiarity" as an alternative means of eliciting the less frequent content bearing (utility or available) words required for balanced vocabulary development as alternative to the "Centers of Interest" approach used in Fundamental French. However, the subjective scaling technique itself appears to be bounded by groups of individuals of like social, cultural, and intellectual levels. Shapiro (1967) demonstrated to his own satisfaction that relative word frequency is a "prothetic" variable and that "magnitude estimation" is a suitable scaling technique for subjective estimation along that continuum. If this be true, we may be able to avoid having to use the large scale objective frequency and range counts we have used in the past by proceding via subjective scaling based on words selected to obtain results equivalent to, or better than, those obtainable from objective counts.

It still is not apparent whether the Richards and Shapiro techniques, if fully developed and proven, will eliminate the deficiencies Bull found in frequency counts, or whether the modifications introduced by Pfeffer in his study of German (oral-topical-empirical approach) will do so, but certainly we should continue to explore them all in an effort to improve our ability to develop better vocabularies for efficient and economical language instruction.

Summary

In summary, it may be said that word frequency counting has evolved complexly in the past 2000 years. With increased knowledge in the physiological, psychological, educational, and linguistic fields, and with the aid of tape recorders and computers we can now do much that we formerly could not. However, much remains to be done in understanding the interrelationship of culture and linguistics; of la langue, and la parole; of the relationships between active and passive vocabularies; and between oral and written language, as well as how best to present them to the student to facilitate his learning. Much also needs to be done in perfecting techniques of language analysis, in order to ensure uniformity of method so that information gained may be better transferred to that common fund of linguistic and cultural knowledge from which future advances may come.

Analyses of the Statistical Lawfulness of Vocabulary Distributions

SECTION III

Language like other natural systems has been an object of study since man has engaged in such enterprises, or at least for as long as we have preserved record of such study. As a natural phenomenon, it presents a uniquely different challenge to the naturalist, however, which is not shared by those systems which can be construed as purely physical. As with other aspects of human behaviour, it preeminently involves intentional motivations which underlie and give purpose to the objective manifestations which are open to study. Thus, the early studies of language as system concentrated almost exclusively upon its intentional aspects; the meanings and symbol processes in whose service it was employed. Two developments, however, presaged a different but parallel method of investigation; the invention of movable type and the rise of enumeration as a measurement tool.

It is the original invention of writing which in very large measure has defined the word entities for which the modern scientist has sought laws. The definition of this entity has remained moot since scribes have sought to record the continous stream of sound which is language. But, with the invention of movable type and the consequent wide distribution of printed language, the definitions employed by the makers of books if nonetheless arbitrary became at least conventional and consistent. For without such consistency their products could not have been successful. Thus, in a sense the problem which this paper seeks to address is a man made problem. We invented a unit called the word for largely commercial purposes and then decided that we should study our own invention by application of another of our inventions, namely counting. Once set in motion, however, the process appears to have assumed a life of its own -- in all regards words appear to have a natural life which share the characteristics of those systems we did not create and their counting has become a scholarly discipline of its own commercial and intrinsic value.

Although measurement by enumeration itself stretches far back into man's time, its early uses were more linguistic and qualitative than quantitative. Measurements of sacks of grain, wealth or live-stock required only that the measurement scale enumerate the finite and directly countable. Such scales have the characteristic that they isomorphically map the objects of enumeration explicitly to an only nominally representative set of numbers. The nominal use of numbers as a measurement device is exemplified by such modern devices as numbering the members of a football team or labeling our coinage with denominations as qualitative categories which only partially reflect their extrinsic values. In such measurement, one moves from few to some through many too many counts. One speaks of a lot of money or more money than can be counted. The enumeration remains limited by the mechanics of physically mapping the objects into their numerical representations. A clay tablet which is to be used to record the number of animals involved in a business transaction serves only because its size and the number of potential mappings are well suited. Notions of an infinite number of animals or of negative amounts of wealth were as meaningless as they were impractical. It is meaningless to declare that I have two and a half cents in my pocket or that if you have five cents in yours that you are twice as wealthy as I. And although you may declare that I owe you three cents in exchange for an article set at that value when I protest that I have zero wealth; i.e., that I have -3 elemental units of money, having minus a billion maximal units of enumeration would be treated in precisely the same way, that is, as without meaning. Such vagaries are inconsistent with the precision which is required of enumeration as a measurement tool. The post Renaissance development and acceptance, however, of arithmetic manipulations which bore no extrinsic relationship to the practical usefulness of enumeration suddenly opened a fertile field of speculative and theoretical implications of the natural lawfulness of the countable. It was not, surprisingly enough given the modern acceptance of such operations, until the 16th century that such arithmetic operations as 3-7=-4 were accepted as other than an absurdity. And only still more recently with the introduction of the Calculus that the succession rule defining infinity has been accepted.

A Taxonomy of Scaling Operations

The process of assigning numbers to phenomena within the structures of a well formulated and explicit set of rules is known as measurement. These measurements, in turn, purport to be the quantification of a defined set of attributes. The measurements represent a model of the attributes which may or may not fit the facts; i.e., may or may not accurately or entirely depict the behavior of the phenomenon in question. One may adjudge the adequacy of the model as representation of the phenomenon it describes either or simultaneously by reference to the accuracy of the deductions derived from the model with respect to the phenomenon's behavior or with respect to the validity of the measurement rules used to derive that model. Thus, much of this paper will be concerned with an evaluation of the adequacy of the fit of variously proposed models of language enumeration and the assumptions of measurement implicit to these models. In order, however, to understand the deeper issues involved in the tests of adequacy of these models, it is necessary first to discuss the broadest implications of measurement per se.

There are four fundemental types of measurement. These forms of measurement differ in the nature and number of assumptions which are held to be characteristic of the qualities they seek to describe.

Nominal scales. The first of these forms of measurement, already alluded to in the opening discussion, assumes only that it is possible to identify the equality or non-equality of any two attributes. Let us take as instance the quality of "wordness." Nominal scaling of this attribute requires only that we correctly assign our units of measurement such that instances of different such qualities receive different measurements and that identities of the quality receive unique measurements. That is to say, we are only required to make explicit those operations which allow us to identify the sameness of the phenomenon to be measured---to recognize the re-occurrence of the same quality attribute when it re-appears and to distinguish such re-appearance from instances which are not the same. Thus, for example, the text of this paper could be re-expressed as a nominal scale which assigned numbers to each of the groups of inkmarks bounded by absence of inkmarks on the basis of their unique patterns. Hence the list:

Language	1
like	2
systems	5
engaged	15
language	92

satisfies the assumptions of a nominal scale in that the numbers do nothing more than uniquely reassign names to the inkmark patterns on the basis of their qualitative characteristics. Under the measurement assumptions of this example, it is inkmark pattern whose equality or non-equality is at issue. If the attribute of "wordness" with respect to some other quality of inkmarks is at question, then we should be required to provide an explicit statement of the recognition of the equality of that aspect of inkiness. Observe that the essential character of the scaling operation remains unchanged if we reassign our measurement numbers by any arbitrary schema so long as we preserve the assumed characteristic of pattern uniqueness. It is, however, a gross violation of that assumption to attribute additional meaning to such a scale. We are not, for example, permitted to assume that the above list of numbers implies that some inkmarks are "larger" or "bigger" than others, and certainly not that some inkmark is X times "larger" or "bigger" than some other identified inkmark. Although the inkmark pattern engaged has been assigned a value which is three times larger in magnitude than that for systems, nothing other than uniqueness of pattern is implied by those assignments. Under the operational assumptions of this measurement operation, we are not permitted to question the values which may have been chosen. However, we may, indeed, question the validity of the assumption that either the quality of pattern uniqueness was correctly identified or that even if correctly identified it has anything meaningful to say about the nature and uses of inkmarks.

Ordinal scales. If we wish to have our measurements reflect the additional attribute of magnitude in its simplest form; i.e., attributions of greater or less than, we are required to make explicit the measurement operations which are to be employed in identifying that attribute. Thus, we might, for example, define the quality of length of inkmark pattern as the measurement operation of comparing each pattern with every other pattern to arrive at judgements of which patterns were longer, shorter or equal to which other patterns. Such a measurement scale for the same example of inkmarks might take the following form.

Language	1
like	8
systems	2
engaged	2
language	1

Observe that this new scale not only identifies inkmarks which are same or unique with respect to pattern as defined by length but additionally quantifies the attribute of "length." It still, however, explicitly does not capture the quality of magnitude of length as it is normally conceived. Hence, the example tendentiously shows the pattern *like* as having a scale value 7 units greater than the pattern with lowest value despite the fact that there are only five patterns. Note, as well that we seem to have serendipitiously captured a quality attribute we had not set out to measure.

Unlike the nominal scale, this scale has assigned the two instances of the "word" *language* the same value. Thus, we can say that the measurement operation defining pattern length apparently more closely matches some of our uses of "words" than does that of ink patterns. It, nonetheless, of course, fails to differentiate items which are clearly differentially used and as such still does not capture any semantic quality.

Interval scales. If we were to require that our measurement scale express the additional attribute of magnitude of difference between patterns, we should have to define an operation by which we assessed that attribute in addition to the definitions already adopted for the other attribute qualities. Thus, we might define pattern "length" as number of discrete inkmarks within a pattern. Such a measurement operation defines "length" strictly in terms of number of elements. Thus, the patterns of the example might be scaled as follows:

Language	8
like	4
systems	7
engaged	2
language	8

Note that now we are permitted to make comparisons of both the attribute qualities of "more than" *and* by how much. Thus, the pattern *language* occupies a magnitude position with respect to *like* which is identical to that of *Language*. It is still not possible, however, under these measurement assumptions to identify the equality of ratios of such magnitude. Thus, we cannot assert that the difference between *language* and *like* stands in the same ratio as Language does to like. In order to make such an ascertainment we would have been required to define the notion of zero magnitude of the quality being assessed. The definition which was provided makes zero length of letters an absurdity or at least unmeasureable. It would be impossible to identify a non-occurrence of a discrete inkmark bounded by non-occurrences of discrete inkmarks. So long as the scale orgin is either undefined or arbitrary with respect to the quality involved in the measurement, we are permitted to transform our measurements to any new set of values which can be expressed as a linear equation of each other. Hence, we are permitted to transform the example values by multiplication of 2 and addition of ten to arrive at the new values X' resulting from the equation: $X' = 2X + 10$.

	X	X'
Language	8	26
like	4	18
systems	7	24
engaged	2	14
language	8	26

These new values of X exhibit precisely the same measurement attributes as the old. The magnitude of the intervals separating each measurement entity has not changed their relative positions with respect to each other.

Ratio scales. The final and most restrictive form of scaling seeks to identify the attribute of equal ratios of quality attributes. In order to do so, such a scale must define the attribute of absolute absence of the quality, equality of attributes, magnitude of difference between attributes and equal ratios of those attributes. Few or none of the scaling techniques typically employed in the social sciences can boast ratio character.

Physics, on the other hand, typically employs such measurement. What is deceiving is the fact that the measurement of numerousity is invariably ratio in form. As a consequence, it is a common error to assume that any set of numbers which can be construed as reflecting the number of some quality is ratio in character. But unless the specific measurement operations which define the necessary assumptions of such a scale are made explicit, the conclusion will certainly lead to misuse of the scale. Thus, for example, if the I.Q. scale is interpreted as a scale of numerousity of intelligence points, we are lead to the gross error of ratio assertions regarding the differences between persons with differing I.Q., not to speak of absolute magnitude assumptions about those differences. Similarly, and more pertinently, measurement of the frequency of word units in a text leaves undefined the attribute of absolute zero occurrence of a unit. Zero frequency is an arbitrarily assigned measurement which ambiguously implies either non-occurrence in the sample or non-occurrence in the population which represents the total language. As a measurement of non-occurrence in the population its real value as a measurement indifferently extends from minus infinity to zero.

The Word as An Attribute of Measurement

Any discussion of measurement as specifically applied to vocabulary must grapple with the definition of the attribute of "wordness." Although it is clear that the user of a language finds the notion of word psychologically meaningful, attempts to make the notion linguistically explicit have not been successful. Greenberg (1957, p. 27) has summarized the linguists position on this matter as follows: "Some linguists deny any validity to the word as a unit, relegating it to folk linguistics. Others believe that the word must be defined separately for each language and that there are probably some languages to which the concept is inapplicable." Nonetheless, Sapir (see Ulman, 1962, p. 39) has observed that "The naive Indian, quite unaccustomed to the concept of the written word, has nevertheless no serious difficulty in dictating a text to a linguist student word by word; he tends, of course, to run his words together as in actual speech, but if he is called to halt and is made to understand what is desired, he can readily isolate the words as such, repeating them as units." For those languages in which there is a rich cultural tradition of writing and literacy, the word as apprehended by its speakers might be construed as little more than the propogation of the conventions of writing. It is in this sense that most counts which define the word as that which is conventionally bounded by spaces in printing define their measurement units. Even in this narrowest of senses, the study of such conventions might be of interest. But, it is the search for the word more broadly considered which is of particular interest: its psychological and linguistic significance.

The word as psychological unit. The child's original exposure to language is solely vocal in form, we reserve instruction in writing and reading until rather late in the child's development, or at least until the spoken language is reasonably in hand. But even the child's earliest vocal experience involves considerable emphasis on those isolatable units of speech which have unitary symbolic value. The child seeks and is given names of things and these names are typically those conventionalized units which we normally call words. In those high cultures for which literacy is sanctified, the parents of children anticipate and transmit the written language conventions. Thus, it is the rare parent of a high culture child who would respond to a query regarding a drum, with the response "Thatsthethingthatgoesboom." It is much more likely that the parent will pare the response down to the minimally isolatable unit of semantic intent which comes closest to the conventional lexical entry for that object: i.e., "drum" accompanied by an appropriate pointing gesture rather than "Thatscalledadrum" or even "Thatsadrum." Further, once the child is made literate, what may have begun as a printer's convention is perceived as a psychological neccessity which takes on its own significance. Later on should this now literate child be required to learn a second language, he will find it both efficacious and satisfying to learn a vocabulary of words for that language and even to expand his own tongue by study of its lexicon. Finally, the adult speaker of a language with written traditions will unerringly identify upon request what is or isn't a word. And even, according to Greenberg, those adults without writing can do the same.

The word as linguistic unit. Assuming then that a reasonable case can be made for the word as a psychological reality, there remains the question as to whether or not there exists a linguistic definition which can serve as an explication of the concept. That is to say, can we provide an explicit theory of "wordness" which is independent of the user's perceptions

BEST COPY AVAILABLE

of the conventions he employs. Such an explication implies what Chomsky (1957) has called "explanatory adequacy" in contrast to the descriptive adequacy which might be served by reference to the conventions of a particular language, whether written or spoken. Accordingly, it is clear that when considered in this light, the answer to such a question lies at the heart of the complete theory of any language and as such will be extraordinarily difficult to attain. The most modern of grammatical treatments which would seek an account of the structure of language typically eschew the problem as premature, choosing instead to assume the weaker requirement embodied in the presumption of a commonsensical appreciation of what a word is as commonly understood (i.e., psychologically apprehended) by the users of the language.

It is thus not accident that one must search backward into the Bloomfieldian era to find attempts at a linguistic definition of the word, an era for which descriptive adequacy was the prime consideration. Bloomfield (1933) attempted to define the word by reference to the formal characteristics of syntactic boundedness. Those minimal forms which can occur as sentences he termed free forms and those which are never used as sentences bound forms. Words, as they are commonly used, are those minimal utterance units which can occur as free forms. What distinguishes words from other free forms is that words cannot be split into still smaller forms without leaving a bound form residue.

It is not difficult to find instances of what the speaker of this language would psychological call words which would not be called words by Bloomfield's definition. All compound forms composed of two or more independent words (by either the conventional definition or the definition under test) such as *penknife* or *yardstick* provide paradoxical exceptions to the definition. Similarly, the functors such as *a* or *the* must occur as bound forms under the definition and yet they are clearly apprehended as psychologically defined words. The meta-language arguments cannot serve to rescue the definition, for all such arguments neccessarily involve the definition we seek as a presumption. Thus, to say that "The." is a permissable sentential response to the question "What is the third word of this sentence?" would only make the issue more cloudly than it already is.

The word as lexical entry. Lexicography at its best represents the structural and functional characteristics of a language as it is conventionally employed, at least, by those who are largely responsible for shaping the culture defined by that language. At its worst, it represents a set of normative prescriptions regarding its language hardly even characterizing its use by those pedants who would prefer proscription to description. The conventionality of either the description or prescription of its source books is largely dictated by the vicissitudes of publishing and data collection. But such conventionality serves, nonetheless, to represent the conventions of the language usage and as a normative model of such usage to itself perpetuate those conventions. The conventions, in turn, capture the aggregate distillation of the psychological realities by which the language user accounts for his language. New words and new usages replace old conventions at the leisurely pace of slow moving publishers who thus assure that the changes have already been accepted as conventions by the majority of their users. All of these factors in combination serve to make the lexicographer's source book an unequaled arbiter of the problems of defining wordness.

The word as grammatical form. Conventionality in language usage extends beyond the boundries of wordness and arbitrary meaning to function and structure. Grammatical classes or parts of speech as they are more traditionally called, codify by label the functional elements which the language user deems essential to his account of the structures he employs. Whether or not such labels have real explanatory meaning in the theory of language is moot. But, again, as conventions they do have at least psychological meaning which even if without linguistic validity at least deserve recognition by dint of the universality of their acceptance in instruction and perception. And, as before, such purposes are best served by the conventionality of the language's dictionary or alternatively as in this research by a structural definition derived from the mutual substitutability of speech parts in language frames which model their usage.

A special set of problems. There exists a grey area of wordness for which no solutions are readily available. Compound forms that have not as yet made the complete and preferred transition from multiple words through hyphenated forms to single units or fixed collocations too extensive in length to move into the hyphenated life form but nonetheless function as if they were single units, and learned forms which because of the pedantry of their users cannot be tolerated to change, all represent exceptional cases for which it is difficult to devise other than ad hoc and arbitrary solutions.

Inflectional forms in those languages for which such grammatical mechanisms are productive do not, however, represent a particularly difficult problem. It is possible to identify variant forms of the simpler root form on the basis of their derivation from a paradigm. Such a paradigm has the characteristics of regularness and of limiting the number of variants to an absolutely small number. Adverbs, in English, for example, are very largely paradigmatically derived from their more productive adjectival roots by the single pattern form of -ly.

A functor may be defined as any free standing word form in analytic languages which is lexically defined as serving strictly grammatical rather than referential functions and for inflectional languages as that morphological change of the stem which carries such meaning. This definition facilitates the counting of both lexical forms and grammatical patterns. In the first instance, the working definition of functor is used to suppress those elements which, occurring with such overwhelmingly high frequency, tend to usurp the lower-frequency, but higher information-content forms. In this sense, "functor' is a convenient catch-all for those terms in a language which are finite in number, but which account for a greatly disproportionate frequency of occurrence. Display 1 illustrates the frequency count equivalent to their total occurrence in the elicited samples. Words which can be generated paradigmatically from a base form can be collapsed into the base form which will then receive a frequency count equivalent to the total occurrence of the paradigm membership; thus, all variations of verbs due to inflections for person, number, and tense can be counted as instances of the base form.

A final word about word. In the end, the final definition of wordness rests entirely upon the conventions of usage in two senses of use. First, we may interpret and operationalize the psychological apperceptions of the language user for an answer to the meaning of word. We require only that the user recognize and distinguish those units which he would construe as words. We do not require that the user explicitly define or understand the processes by which such recognition is achieved. Where dictionaries exist, these source books provide the best aggregate judgements of such recognition, where they do not we shall have to compile such judgements directly from the speakers themselves. In the second sense of use, it is the purposes of our definition of wordness which must be examined. In this paper we shall be focusing on the statistical lawfulness of word occurrences. The test of alternative definitions of the word as unit of measurement rests entirely upon the empirical comparisons of the outcomes of these definitions. Does it as matter of empirical fact, made a difference in the characteristics of the functional lawfulness of vocabulary to define root variants as separate or same forms? When the uses of our definition of word are pedogical rather than theoretical, it is surely certain that we shall at least require other tests of that definition; tests which will involve considerations which are as practical as the model tests are theoretical.

Statistics and Measurement: The Schemapiric View

Before beginning our survey of the statistical models which have been proposed for the distribution of vocabulary in language, it is appropriate to forewarn the reader of the theoretical distinction between counting and modeling, between empirics and schematics. Since that distinction has been for some time the special concern of S.S. Stevens, it is appropriate to quote him on the schemapiric principle at some length. "Although measurement began in the empirical mode, with the accent on the counting of moons and paces and warriors, it was destined in modern times to find itself debated in the formal, schematic, syntactical mode, where models can be made to bristle with symbols. Mathematics, which like logic constitutes a formal endeavor, was not always regarded as an arbitrary construction devoid of substantive content, an adventure of postulate and theorem. In early ages mathematics and empirical measurement were as warp and woof, interpenetrating each other so closely that our ancestors thought it proper to prove arithmetic theorems by resort to counting or to some other act of measurement. The divorce took place only in recent times. And mathematics now enjoys full freedom to 'play upon symbols,' as Gauss phrased it, with no constraints imposed by the demands of empirical measurement.

"So also with other formal or schematic systems. The propositions of a formal logic express tautologies that say nothing about the world of tangible stuff. They are analytic statements, so-called, and they stand apart from the synthetic statements that express facts and relations among empirical objects. There is a useful distinction to be made between the analytic, formal, syntactical propositions of logic and the synthetic, empirical statements of substantive discourse.

"Probability exhibits the same double aspect, the same schemapiric nature. Mathematical theories of probability inhabit the formal realm as analytic, tautologous, schematic systems, and they say nothing at all about dice, roulette, or lotteries. On the empirical level, however, we count and tabulate events at the gaming table or in the laboratory and note their relative frequencies. Sometimes the relative frequencies stand in isomorphic relation to some property of a mathematical model of probability; at other times the observed frequencies exhibit scant accord with 'expectations.'" (S.S. Stevens, 1968.)

It is obvious that Stevens might as readily and appropriately have cited the counting of words in the above passage. Adopting this schemapiric point of view, we shall for each of the models of vocabulary distribution to be reviewed, separately examine the schematic assumptions of the models, their fit to the empirical data and the psychological justification of those assumptions. But before proceeding there is still another consideration which must be addressed by any statistical model designed to account for an empirical domain, namely, the methodological problems of sampling.

Methodological issues in sampling. If one wishes to construe a selected corpus of language to be representative of some larger body of language of which that corpus is sample, the researcher is compelled to provide a rational defence of the sample's representativeness. Selection by random strategy is designed to provide such justification on the grounds that a random sample requires that all members of the population had equal probability of being selected as members of that sample. Under such rationale, the occurrence frequencies of the units of analysis are both efficient and unbiased estimators of the population probabilities of those units. But then two problems arise, random with respect to what and how are we to translate random into a set of explicit procedures? The overwhelming bulk of research on vocabulary has concentrated on the written forms of language, the number of worthwhile spoken analyses numbers less than half a dozen. The preceding sections have reviewed and evaluated these studies. The populations represented by the spoken and written forms of a language are both different and same when viewed from differing standpoints. We have argued that at the level of the functor, the vocabularies of speech and writing are as alike as the linguistic code is inflexible with respect to their grammatical function. At the level of substantive choices, the two are as separate as the distinction made by the culture between informal and formal styles of communication, with an extensive penumbra area of overlap between those styles at the level of the higher frequency substantives. And from still another viewpoint, the two communication forms may or may not be different with respect to their schemapiric lawfulness, a consideration which we are now deferring.

But in a sense even the distinction being made between speech and writing is itself artificial for some purposes. Plays are written to be spoken and all writing must be speakable if it is to conform to its parent linguistic code. Nor is it simple to classify the procedure which this research proposes as the optimal sampling strategy; that of eliciting restricted associations from the users of a language. That procedure is designed to bypass the written-spoken dichotomy by sampling from the highest frequency items of the users vocabulary. The rationale of that assumption, in turn, rests upon the spew hypothesis. Under that rationale the problem of corpus length is also largely avoided, for no attempt is being made to fully sample the entire frequency range of vocabulary items as they appear in the population. The spew hypothesis quite simply posits that "...the order of emission of verbal units is directly related to frequency of experience with those units." (Underwood and Schulz, 1960.)

A number of studies have provided strong support for such an assertion. Johnson (1956) demonstrated that 84% of the most frequent associations to the Kent-Rosanoff stimuli occured with a frequency of 50 times or more per million in the Thorndike-Lorge list, whereas only 48% of the least frequent responses had equally high ratings. Howes (1957) computed the correlation between frequency of associations to the Kent-Rosanoff list and frequency of words in the language to be .94 if functors are excluded from consideration. The effect has even been demonstrated when subjects are asked to provide male given names; those names which occur most frequently in the written language are also those most likely to be given by a subject (Cromwell, 1956). Bousfield and Barclay (1950) have also demonstrated that the order of emission of verbal units is directly correlated with their frequency of occurrence in the language.

Taken in its weakest sense, the spew hypothesis is not an hypothesis at all. It is obvious that if emission of verbal units is taken to include all uses of the language, the complete tabulations of such emissions are the frequencies of those units. But in its strongest sense, the spew hypothesis provides a sampling strategy for estimating the total linguistic probability of verbal units. Construed as ad libitum responses, associational responses obtained from subjects provide a higher face validity procedure for estimating the frequency of spoken language units.

Either spoken or written data suffer from several inherent difficulties which accure to the nature of natural language codes. The lawful statistical nature of such counts always produces a frequency ordering in which roughly half of the occurrence types have token realizations which are at the limits of measurement: i.e., have single occurrence frequencies. Probability estimates of population frequencies from such inherently errorful sample frequencies are statistically unreliable. At the high frequency end of the distribution of such word samples one consistently finds that function and interstitial words account for disproportionately high percentages of the total sample. The situation is roughly analogous to using the Wall Street Journal to determine the frequency of English units. From such a data base, ordinal numbers and fractions would dominate the frequency distribution of this disparity. In English, nouns, adjectives, verbs, and "pure" adverbs comprise over 99 percent of the total available vocabulary presented in the Shorter Oxford English Dictionary; in contrast to this, we have for all the remaining parts of speech not more than 650 words. Yet these two groups provide approximately equal proportions of the total word usage. While all Group II words in English are not strictly "functors", they all share three features of functors: (1) they belong to a small, limited, isolatable class; (2) they have paradigmatic features; (3) they occur with extremely high frequency and, thus, suppress non-functorlike Group I words. It is, therefore, our contention that functorlike words should be treated separately, both for lexical counts and, as it turns out, for grammatical pattern counts.

In the case of strictly inflectional languages, the paradigmatic functors will occur as bound forms in traditional orthography. This presents no problem other than identifying these forms and coding them in such a manner that the "root" form will be the entry into the frequency count. In Latin, for example, *agricolae* would be subsumed into *agricola*.

Approximate Occurrences of Parts of Speech in Shorter Oxford English Dictionary

Group I

Nouns	58,000
Adjectives	27,000
Verbs	13,500
Adverbs(*)	150
	98,650 (approx. total)

Group II

Pronouns	100
Prepositions	100
Conjunctions	50
Aux. Verbs	10
Articles	2
	262 (approx. total)

(*) Counts only "pure" adverbs not derived paradigmatically from adjectives

Display I. Estimates of vocabulary words in different parts of speech available in the English language (Yule, 1944).

If the text contained one instance of *agricolae* and one of *agricola*, the frequency tally would show *agricola* as occurring twice.

However, pure analytic and pure inflectional languages are the exception, not the rule. Therefore, the treatment of "functors" in the hybrid languages must allow for the uncluttered tally of words, yet preserve the grammatical patterning of occurrences. Thus, in German, for example, *zum kleinen Kind* would be coded as preposition-definite article-adjective-noun for grammatical pattern and *klein* would be tabulated in its base form for frequency tally. Similarly, in English, *cat* and *cats* would appear as two occurrences of *cat*, since, in English the two forms can be considered as co-occurring items of a paradigm. Verbs would be treated similarly for frequency counts. The total tally for the verb, *run*, for example, would include occurrences of paradigmatic forms such as *runs*, *ran*, and *running*.

There are other common words which should be given separate treatment. For example, numbers, certain kinship terms, days of the week, month of the year, and the like require special attention. The term *Monday* should be taken to include the terms for the other days of the week as though it were a root form from which the others are derived. Thus, all names for the days of the week which are elicited would contribute to the frequency total for the base form, arbitrarily taken to be *Monday*. Similarly, in English, the terms for the members of the nuclear family (*father, mother, son, daughter, brother, sister, husband, wife*) should share a position in the count. Functors, as is the case with numbers, are important to the language, but they displace and minimize the importance of the other substantive form classes because of their overwhelming prominence in natural languages. Foreign language instruction has typically met this difficulty by sub-dividing the lexical units of the language into separate form classes. Such form classes are fundamental to any description of a language. They function at elemental levels in both phrase structure and transformational rules. The speaker of a language only rarely can make explicit the category rules which define such grammatical classes and, even in these rare cases, such explicitness is typically incorrect. However, the speaker does use such rules in the construction of any utterance, his inability to provide an explicit account of the nature of those rules is not evidence against their functional utility. If the speaker is given a contextual frame which calls for a unit from a particular grammatical class, the speaker can provide an appropriate completion. Further, the choice of the particular completion within that functional class is apparently determined by the frequency of experience of that unit. Thus, elicitation procedures which call for grammatical class associations in specified frames simultaneously solve two problems otherwise encountered in frequency counts: 1) all token frequencies are automatically marked by function class and 2) frequency determinations of unit types are separately determined within function class, thus increasing the pay-off yield of the data collection.

For the models of continuous language samples to be reviewed, the issue of corpus length is as crucial as it is difficult to answer. Rapoport

(1965) addressing himself to this problem with regard to speech samples has argued that:

> "In the selection of speech to be analyzed, the question of how long the transcript should be, though practically important, is not easy to answer. Intuitively it would be nice to have very long transcripts, 5000 words or more, in order to get a substantial sample of the subject's vocabulary. Practical considerations, on the other hand, call for smaller samples. In addition, it might not be feasible to obtain very long samples of connected discourse from the subject. Without considering some exceptions, people usually do not utter 5000 words and more in one session on the same topic. It seems that a proper solution to the length of the transcript is an empirical one. Sample sizes should be considered within the range where the mathematical form of the observed distribution of word-frequencies is not markedly changed."

And then after reviewing data similarly collected by Howes and Geschwind (1962) who claimed that: "These data show that even for samples of 1000 words, there is excellent correspondence between the theoretical equation and the empirical distributions. The considerations suggest that for most purposes samples of 2000 words are adequate for estimating parameters of [spoken] word-frequency.", Rapoport concludes that: "It thus seems that the sample sizes used [in the Rapoport study] (between 1000 and 5000 tokens) are appropriate."

BEST COPY AVAILABLE

Word Frequency Comparisons Among First 400 Entries of Three Counts

	Kučera-Francis		American-Heritage		Black-Ausherman	
	TYPE	F	TYPE	F	TYPE	F
1	THE	69971	THE	373123	THE	15067
2	OF	36411	OF	146201	OF	4910
3	AND	28852	AND	133899	AND	10661
4	TO	26149	TO	121347	TO	9237
5	A	23237	A	124959	A	7436
6	IN	21341	IN	99108	IN	6188
7	THAT	10595	THAT	47443	THAT	5627
8	IS	10099	IS	60852	IS	4982
9	WAS	9816	WAS	40534	WAS	2373
10	HE	9543	HE	46249	HE	2233
11	FOR	9489	FOR	39322	FOR	2140
12	IT	8756	IT	47204	IT	4458
13	WITH	7289	WITH	30455	WITH	1479
14	AS	7250	AS	32208	AS	1545
15	HIS	6997	HIS	29268	HIS	[illegible]
16	ON	6742	ON	36482	ON	1962
17	BE	6377	BE	23746	BE	2301
18	AT	5378	AT	23975	AT	1289
19	BY	5305	BY	20189	BY	1240
20	I	5173	I	25932	I	1949
21	THIS	5146	THIS	23301	THIS	2974
22	HAD	5133	HAD	20511	HAD	1063
23	NOT	4609	NOT	18645	NOT	1240
24	ARE	4393	ARE	35454	ARE	2375
25	BUT	4381	BUT	19196	BUT	1557
26	FROM	4369	FROM	22799	FROM	1063
27	OR	4207	OR	21283	OR	1576
28	HAVE	3941	HAVE	22337	HAVE	2611
29	AN	3747	AN	14696	AN	913
30	THEY	3618	THEY	27620	THEY	3063
31	WHICH	3562	WHICH	14016	WHICH	1262
32	ONE	3292	ONE	19976	ONE	1625
33	YOU	3286	YOU	50957	YOU	3195
34	WERE	3284	WERE	17031	WERE	1087
35	HER	3037	HER	11375	HER	144
36	ALL	3001	ALL	16997	ALL	1522
37	SHE	2859	SHE	13653	SHE	165
38	THERE	2724	THERE	15194	THERE	1662
39	WOULD	2714	WOULD	11188	WOULD	904
40	THEIR	2670	THEIR	13258	THEIR	855
41	WE	2653	WE	16452	WE	2830
42	HIM	2619	HIM	10703	HIM	360
43	BEEN	2472	BEEN	7645	BEEN	808
44	HAS	2439	HAS	10369	HAS	1447
45	WHEN	2331	WHEN	15886	WHEN	959
46	WHO	2252	WHO	7576	WHO	701
47	WILL	2244	WILL	12646	WILL	1571
48	MORE	2216	MORE	9992	MORE	636
49	NO	2201	NO	8483	NO	447
50	IF	2199	IF	12907	IF	960
51	OUT	2096	OUT	12252	OUT	960
52	SO	1984	SO	11543	SO	926
53	SAID	1961	SAID	15309	SAID	240
54	WHAT	1908	WHAT	17709	WHAT	998
55	UP	1895	UP	12776	UP	983
56	ITS	1858	ITS	7512	ITS	240
57	ABOUT	1815	ABOUT	12496	ABOUT	981
58	INTO	1791	INTO	10620	INTO	616
59	THAN	1709	THAN	7982	THAN	412
60	THEM	1789	THEM	11997	THEM	895
61	CAN	1772	CAN	15247	CAN	928
62	ONLY	1747	ONLY	6583	ONLY	472
63	OTHER	1702	OTHER	10729	OTHER	761
64	NEW	1635	NEW	5448	NEW	308
65	SOME	1617	SOME	11524	SOME	812
66	COULD	1599	COULD	8585	COULD	439
67	TIME	1599	TIME	8441	TIME	906
68	THESE	1573	THESE	11611	THESE	885
69	TWO	1412	TWO	10085	TWO	114
70	MAY	1400	MAY	6635	MAY	305
71	THEN	1377	THEN	12022	THEN	643
72	DO	1363	DO	12695	DO	843
73	FIRST	1360	FIRST	7655	FIRST	593
74	ANY	1345	ANY	5023	ANY	520
75	MY	1319	MY	7898	MY	315
76	NOW	1314	NOW	7457	NOW	997
77	SUCH	1303	SUCH	4223	SUCH	303
78	LIKE	1290	LIKE	9696	LIKE	632
79	OUR	1252	OUR	5777	OUR	1271
80	OVER	1236	OVER	6882	OVER	738
81	MAN	1207	MAN	5486	MAN	502
82	ME	1181	ME	6186	ME	188
83	EVEN	1171	EVEN	4225	EVEN	235
84	MOST	1160	MOST	5785	MOST	513
85	MADE	1125	MADE	7073	MADE	335
86	AFTER	1070	AFTER	5915	AFTER	570
87	ALSO	1069	ALSO	4647	ALSO	524
88	DID	1044	DID	7165	DID	265
89	MANY	1030	MANY	12158	MANY	653
90	BEFORE	1016	BEFORE	5275	BEFORE	360
91	MUST	1013	MUST	4307	MUST	419
92	THROUGH	969	THROUGH	5442	THROUGH	382
93	BACK	967	BACK	5862	BACK	477
94	YEARS	949	YEARS	3986		0
95	WHERE	938	WHERE	5611	WHERE	313
	MUCH	937	MUCH	5386	MUCH	445
	YOUR	923	YOUR	15311	YOUR	649

BEST COPY AVAILABLE

98	MAY	909	MAY	6635	MAY	365
99	WELL	897	WELL	4255	WELL	744
100	DOWN	895	DOWN	7206	DOWN	434
101	SHOULD	888	SHOULD	3470	SHOULD	400
102	BECAUSE	883	BECAUSE	4307	BECAUSE	641
103	EACH	877	EACH	14231	EACH	215
104	JUST	872	JUST	5858	JUST	744
105	THOSE	850	THOSE	2647	THOSE	[illegible]
106	PEOPLE	847	PEOPLE	7180	PEOPLE	[illegible]
107	MR	839	MR	748		
108	HOW	834	HOW	13301	HOW	406
109	TOO	832	TOO	5071	TOO	[illegible]
110	LITTLE	831	LITTLE	6204	LITTLE	[illegible]
111	STATE	808		0	STATE	176
112	GOOD	807	GOOD	5343	GOOD	521
113	VERY	796	VERY	5997	VERY	[illegible]
114	MAKE	794	MAKE	8333	MAKE	[illegible]
115	WORLD	787	WORLD	2799	WORLD	557
116	STILL	782	STILL	3421	STILL	165
117	OWN	772	OWN	3006	OWN	179
118	SEE	763	SEE	8518	SEE	447
119	MEN	760	MEN	4067	MEN	639
120	WORK	755	WORK	4358	WORK	520
121	LONG	750	LONG	6220	LONG	244
122	GET	735	GET	5700	GET	891
123	HERE	730	HERE	4184	HERE	350
124	BETWEEN	730	BETWEEN	3324	BETWEEN	150
125	BOTH	721	BOTH	2646		0
126	LIFE	715	LIFE	2612	LIFE	239
127	BEING	712	BEING	3092	BEING	246
128	UNDER	707	UNDER	2987	UNDER	188
129	NEVER	698	NEVER	3135	NEVER	143
130	DAY	686	DAY	5018	DAY	410
131	SAME	686	SAME	5022	SAME	240
132	ANOTHER	683	ANOTHER	4377	ANOTHER	337
133	KNOW	683	KNOW	5655	KNOW	769
134	WHILE	680	WHILE	2837	WHILE	178
135	LAST	676	LAST	3330	LAST	277
136	MIGHT	672	MIGHT	2824	MIGHT	229
137	US	672	US	3928	US	408
138	GREAT	665	GREAT	3855	GREAT	563
139	OLD	660	OLD	3894	OLD	189
140	YEAR	660	YEAR	2277	YEAR	590
141	OFF	639	OFF	3873	OFF	297
142	COME	630	COME	4676	COME	545
143	SINCE	628	SINCE	2041	SINCE	161
144	AGAINST	626	AGAINST	1755	AGAINST	130
145	GO	626	GO	5380	GO	1356
146	CAME	622	CAME	4914	CAME	226
147	RIGHT	613	RIGHT	4815	RIGHT	361
148	USED	612	USED	5607		0
149	TAKE	611	TAKE	4089	TAKE	251
150	THREE	610	THREE	4413		0
151	STATES	605		0		0
152	HIMSELF	603	HIMSELF	1789		0
153	FEW	601	FEW	2685	FEW	304
154	HOUSE	591	HOUSE	2705	HOUSE	117
155	USE	589	USE	7009	USE	593
156	DURING	585	DURING	1924	DURING	186
157	WITHOUT	583	WITHOUT	2392	WITHOUT	74
158	AGAIN	578	AGAIN	3892	AGAIN	159
159	PLACE	571	PLACE	4240	PLACE	287
160	AMERICAN	569	AMERICAN	1575		0
161	AROUND	561	AROUND	4632	AROUND	306
162	HOWEVER	552	HOWEVER	1314	HOWEVER	160
163	HOME	547	HOME	3308	HOME	178
164	SMALL	542	SMALL	3555	SMALL	215
165	FOUND	536	FOUND	3362	FOUND	157
166	MRS	534	MRS	1619		0
167	THOUGHT	515	THOUGHT	2635	THOUGHT	182
168	WENT	507	WENT	4132	WENT	253
169	SAY	504	SAY	3916	SAY	465
170	PART	500	PART	4245	PART	294
171	ONCE	499	ONCE	2435	ONCE	95
172	GENERAL	497		0	GENERAL	143
173	HIGH	497	HIGH	2237	HIGH	287
174	I	496				
175	UPON	495	UPON	1574	UPON	172
176	SCHOOL	492	SCHOOL	2599	SCHOOL	277
177	EVERY	491	EVERY	3398	EVERY	314
178	DONT	489	DONT	2881	DONT	369
179	DOES	485	DOES	4408	DOES	156
180	GOT	482	GOT	2626	GOT	250
181	UNITED	482		0		0
182	LEFT	480	LEFT	2885	LEFT	129
183	NUMBER	472	NUMBER	6054	NUMBER	146
184	COURSE	465	COURSE	1352	COURSE	306
185	WAR	464		0	WAR	1134
186	UNTIL	461	UNTIL	2596	UNTIL	[illegible]
187	ALWAYS	458	ALWAYS	2657	ALWAYS	[illegible]
188	AWAY	456	AWAY	3814	AWAY	135
189	SOMETHING	450	SOMETHING	2761	SOMETHING	[illegible]
190	TWO	450	TWO	10085	TWO	114
191	FACT	447		0	FACT	261
192	THOUGH	442	THOUGH	1331		0
193	WATER	442	WATER	7194	WATER	159

BEST COPY AVAILABLE

194	LESS	438	LESS	1360	LESS	96
195	PUBLIC	438		0		0
196	PUT	437	PUT	[illegible]	PUT	342
197	THINK	433	THINK	4746	THINK	559
198	ALMOST	432	ALMOST	[illegible]	ALMOST	113
199	HAND	431	HAND	2316	HAND	178
200	ENOUGH	430	ENOUGH	2763	ENOUGH	141
201	FAR	427	FAR	2250	FAR	142
202	TOOK	426	TOOK	2490	TOOK	[illegible]
203	HEAD	424	HEAD	2487	HEAD	[illegible]
204	YET	419	YET	1325	YET	119
205	GOVERNMENT	417		0	GOVERNMENT	331
206	SYSTEM	416		0	SYSTEM	126
207	BETTER	414	BETTER	[illegible]	BETTER	[illegible]
208	SET	414	SET	[illegible]	SET	196
209	TOLD	413	TOLD	[illegible]		0
210	NOTHING	412	NOTHING	1420		0
211	RIGHT	411	RIGHT	2307	RIGHT	128
212	END	410	END	2181	END	131
213	WHY	404	WHY	4147	WHY	201
214	CALLED	401	CALLED	5784		0
215	DIDNT	401	DIDNT	2016	DIDNT	145
216	EYES	401	EYES	2303		0
217	FIND	399	FIND	6956	FIND	161
218	GOING	399	GOING	2832		0
219	LOOK	399	LOOK	4333	LOOK	[illegible]
220	ASKED	398	ASKED	2926		0
221	LATER	397	LATER	[illegible]		0
222	KNEW	395	KNEW	2044		0
223	POINT	395	POINT	1904	POINT	201
224	NEXT	394	NEXT	2727	NEXT	162
225	PROGRAM	394		0	PROGRAM	96
226	CITY	393	CITY	1843	CITY	165
227	BUSINESS	392		0	BUSINESS	111
228	GIVE	391	GIVE	1365	GIVE	303
229	GROUP	390	GROUP	3570	GROUP	120
230	TOWARD	386	TOWARD	1690		0
231	YOUNG	385	YOUNG	1920	YOUNG	113
232	DAYS	384	DAYS	2003		0
233	LET	384	LET	2176	LET	150
234	ROOM	383	ROOM	1901	ROOM	115
235	PRESIDENT	382		0		0
236	SIDE	380	SIDE	2532	SIDE	144
237	SOCIAL	380		0		0
238	GIVEN	377	GIVEN	1661	GIVEN	132
239	PRESENT	377		0	PRESENT	153
240	SEVERAL	377	SEVERAL	1801	SEVERAL	158
241	ORDER	376	ORDER	1507	ORDER	167
242	NATIONAL	375		0		0
243	POSSIBLE	373		0	POSSIBLE	[illegible]
244	RATHER	373		0	RATHER	150
245	SECOND	373	SECOND	2094	SECOND	[illegible]
246	FACE	371	FACE	1629		0
247	PER	371		0		0
248	AMONG	370	AMONG	1308		0
249	FORM	370	FORM	2720	FORM	137
250	IMPORTANT	369	IMPORTANT	2598	IMPORTANT	[illegible]
251	OFTEN	369	OFTEN	2611		0
252	THINGS	368	THINGS	4070		0
253	LOOKED	367	LOOKED	3197		0
254	EARLY	366	EARLY	1439		0
255	WHITE	365	WHITE	2085		0
256	CASE	362		0	CASE	146
257	JOHN	362	JOHN	1347		0
258	BECOME	361	BECOME	1304	BECOME	171
259	LARGE	361	LARGE	2777	LARGE	[illegible]
260	BIG	360	BIG	3476	BIG	[illegible]
261	NEED	360	NEED	2081	NEED	108
262	FOUR	359	FOUR	2357		0
263	WITHIN	359		0		0
264	FELT	357		0		0
265	ALONG	355	ALONG	2835	ALONG	157
266	CHILDREN	355	CHILDREN	2575		0
267	SAW	352	SAW	2900		0
268	BEST	351	BEST	1984		0
269	CHURCH	348		0		0
270	EVER	345	EVER	[illegible]	EVER	154
271	LEAST	343		0		0
272	POWER	342		0	POWER	150
273	DEVELOPMENT	334		0		0
274	LIGHT	333	LIGHT	2376	LIGHT	116
275	THING	333	THING	1828	THING	754
276	SEEMED	332		0		0
277	FAMILY	331	FAMILY	1768		0
278	INTEREST	330		0	INTEREST	216
279	WANT	329	WANT	2155	WANT	363
280	MEMBERS	325		0		0
281	MIND	325		0	MIND	116
282	COUNTRY	324	COUNTRY	[illegible]	COUNTRY	[illegible]
283	AREA	323		0		0
284	OTHERS	323	OTHERS	1903		0
285	DONE	320	DONE	1566	DONE	195
286	TURNED	320	TURNED	1785		0
287	ALTHOUGH	319		0		0
288	OPEN	319	OPEN	1516		0
289	GOD	318		0		0

BEST COPY AVAILABLE

290	SERVICE	315		0	SERVICE	98
291	CERTAIN	313	CERTAIN	1198	CERTAIN	225
292	KIND	313	KIND	2262	KIND	105
293	PROBLEM	313		0	PROBLEM	[illegible]
294	BEGAN	312	BEGAN	2491		0
295	DIFFERENT	312	DIFFERENT	3026	DIFFERENT	213
296	DOOR	312	DOOR	1788		0
297	THUS	312		0		0
298	HELP	311	HELP	3175	HELP	108
299	SENSE	311		0		0
300	MEANS	310	MEANS	1962		0
301	WHOLE	309	WHOLE	2291	WHOLE	111
302	MATTER	308		0		0
303	PERHAPS	307	PERHAPS	1235	PERHAPS	103
304	ITSELF	304		0		0
305	ITS	302	ITS	7512	ITS	240
306	YORK	301		0		0
307	TIMES	300	TIMES	2051		0
308	HUMAN	299		0		0
309	LAW	299		0		0
310	LINE	298	LINE	3293	LINE	137
311	ABOVE	296	ABOVE	2298		0
312	NAME	294	NAME	3766	NAME	96
313	EXAMPLE	292	EXAMPLE	1933	EXAMPLE	149
314	ACTION	291		0		0
315	COMPANY	290		0		0
316	HANDS	289	HANDS	1357		0
317	LOCAL	288		0		0
318	SHOW	287	SHOW	2734	SHOW	439
319	FIVE	286	FIVE	1725		0
320	HISTORY	286		0	HISTORY	193
321	WHETHER	286		0	WHETHER	[illegible]
322	GAVE	285	GAVE	1534		0
323	EITHER	284		0		0
324	TODAY	284		0	TODAY	310
325	ACT	283		0		0
326	FEET	283	FEET	2545	FEET	140
327	ACROSS	282	ACROSS	1942		0
328	THREE	282	THREE	4413		0
329	PAST	281		0		0
330	QUITE	281		0	QUITE	240
331	TAKEN	281		0		0
332	ANYTHING	280		0	ANYTHING	138
333	HAVING	279		0		0
334	SEEN	279	SEEN	1663		0
335	DEATH	277		0		0
336	BODY	276	BODY	1783	BODY	31
337	EXPERIENCE	276		0		0
338	HALF	275	HALF	1514	HALF	104
339	REALLY	275	REALLY	1414	REALLY	233
340	WEEK	275		0	WEEK	174
341	CAR	274	CAR	1757	CAR	120
342	FIELD	274		0	FIELD	164
343	WORD	274	WORD	7532	WORD	151
344	WORDS	274	WORDS	[illegible]		0
345	ALREADY	273		0		0
346	THEMSELVES	270		0		0
347	INFORMATION	269		0		0
348	IN	268	IN	1848	IN	95
349	TELL	268	TELL	3715	TELL	183
350	COLLEGE	267		0	COLLEGE	146
351	SHALL	267		0		0
352	TOGETHER	267	TOGETHER	2629	TOGETHER	53
353	MONEY	265	MONEY	1694		0
354	PERIOD	265		0		0
355	HELD	264		0		0
356	KEEP	264	KEEP	2509	KEEP	197
357	SURE	264	SURE	1956	SURE	113
358	PROBABLY	261		0		0
359	FREE	260		0		0
360	REAL	260		0		0
361	SEEMS	259		0		0
362	BEHIND	258	BEHIND	1376		0
363	CANNOT	258	CANNOT	1279		0
364	MISS	258		0		0
365	POLITICAL	258		0		0
366	AIR	257	AIR	3673	AIR	155
367	QUESTION	257		0		0
368	MAKING	255	MAKING	1408		0
369	OFFICE	255		0		0
370	BROUGHT	253	BROUGHT	1357		0
371	WHOSE	252		0		0
372	SPECIAL	250		0		0
373	HEARD	247	HEARD	1788	HEARD	132
374	MAJOR	247		0		0
375	PROBLEMS	247		0		0
376	AGO	246	AGO	1387	AGO	126
377	BECAME	246	BECAME	1501	BECAME	96
378	FEDERAL	246		0		0
379	MOMENT	246		0		0
380	STUDY	246	STUDY	2581	STUDY	33
381	AVAILABLE	245		0		0
382	KNOWN	245	KNOWN	1401		0
383	RESULT	244		0		0
384	STREET	244		0		0
385	ECONOMIC	243		0		0

BEST COPY AVAILABLE

```
386  BOY           242   BOY      2529   BOY        111
387  POSITION      241               0   POSITION    99
388  REASON        241               0   REASON     [illegible]
389  CHANGE        240   CHANGE   1854   CHANGE     [illegible]
390  [illegible]   [illegible]       0                0
391  BOARD         [illegible]       0                0
392  INDIVIDUAL    239               0                0
393  JOB           [illegible]       0   JOB        179
394  SOCIETY       237               0                0
395  AREAS         236               0                0
396  WEST          235               0                0
397  CLOSE         234   CLOSE    1288   CLOSE      112
398  TURN          233   TURN     1577   TURN       152
399  LOVE          232               0                0
400  COMMUNITY     231               0                0
```

Items Not Within First 400 Words of Kučera-Francis Count From First 400 of

BEST COPY AVAILABLE

American-Heritage TYPE	F	Black-Ausherman TYPE	F
WAY	6612	STATE	808
WRITE	9846	STATES	605
READ	3057	GENERAL	497
SOUND	4667	UNITED	482
BELOW	3270	WAR	464
FOOD	2801	FACT	447
LAND	2953	PUBLIC	436
ANIMALS	2625	GOVERNMENT	417
SOMETIMES	2278	SYSTEM	416
LIVE	2431	PROGRAM	394
PAGE	2831	BUSINESS	392
EARTH	2690	PRESIDENT	382
MOTHER	2343	SOCIAL	380
PARTS	2331	PRESENT	377
FATHER	2245	NATIONAL	375
FOLLOWING	2680	POSSIBLE	373
PICTURE	2500	RATHER	373
SOON	2129	PER	371
STORY	2237	CASE	362
BOYS	2155	WITHIN	359
PAPER	2372	FELT	357
HARD	1980	CHURCH	348
NEAR	1985	LEAST	343
SENTENCE	3122	POWER	342
TPDAY	1923	DEVELOPMENT	334
TRY	1958	SEEMED	332
MILES	2146	INTEREST	330
SUN	1977	MEMBERS	325
WAYS	1899	MIND	325
HEAR	2154	AREA	323
ANSWER	2002	ALTHOUGH	319
SEA	1812	GOD	318
TOP	1741	SERVICE	315
LEARN	1674	PROBLEM	313
PLAY	2113	THUS	312
USING	1825	SENSE	311
USUALLY	1712	MATTER	308
MORNING	1736	ITSELF	304
TREES	1645	YORK	301
MOVE	1592	HUMAN	299
CUT	1757	LAW	299
TRUE	1696	ACTION	291
SENTENCES	2611	COMPANY	290
RED	1557	LOCAL	288
FISH	1513	HISTORY	286
PLANTS	1886	WHETHER	286
LIVING	1645	EITHER	284
WANTED	1637	TODAY	284
BLACK	1556	ACT	283
EAT	1616	PAST	281
SHORT	1534	QUITE	281
UNITEDSTA	1836	TAKEN	281
RUN	1473	ANYTHING	280
KINDS	1545	HAVING	279
BOOK	1453	DEATH	277
GROUND	1511	EXPERIENCE	276
LINES	1715	WEEK	275
COLD	1469	FIELD	274
TABLE	1502	ALREADY	273
REMEMBER	1423	THEMSELVES	270
TREE	1421	INFORMATION	269
ZERO	1711	COLLEGE	267
FRONT	1438	SHALL	267
SPACE	1499	PERIOD	265
INSIDE	1398	HELD	264
ILL	1604	PROBABLY	261
LEARNED	1345	FREE	260
STARTED	1409	REAL	260
IDEA	1397	SEEMS	259
CALL	1374	MISS	258
LIVED	1372	POLITICAL	258
MAKES	1311	QUESTION	257
LOOKING	1331	OFFICE	255
ADD	1654	WHOSE	252
GROW	1418	SPECIAL	250
DRAW	1623	MAJOR	247
WIND	1336	PROBLEMS	247
PLACES	1337	FEDERAL	246
LETTER	1738	MOMENT	246
LETTERA	1535	AVAILABLE	245
LETTERS	1798	RESULT	244
COMES	1289	STREET	244
ABLE	1260	ECONOMIC	243
DOG	1380	POSITION	241
SHOWN	1490	REASON	241
MEAN	1266	SOUTH	240
ENGLISH	1564	BOARD	239
REST	1183	INDIVIDUAL	239
SIX	1229	JOB	238
FELL	1275	SOCIETY	237
FIRE	1227	AREAS	236
READY	1207	WEST	235
GREEN	1216	LOVE	232
YES	1317	COMMUNITY	231

BEST COPY AVAILABLE

YEARS	949	WAY	488
MR	839	ARMY	439
BOTH	721	PLANE	375
USED	612	START	367
THREE	610	BELIEVE	299
STATES	605	LIVE	294
HIMSELF	603	FIGHT	275
AMERICAN	569	IDEA	272
MRS	534	IMPROVE	265
UNITED	482	MEAN	254
THOUGH	442	CALL	240
PUBLIC	438	TYPE	236
TOLD	413	PROBABLE	235
NOTHING	412	SHIP	227
CALLED	401	FELLOW	226
EYES	401	PLAN	224
GOING	399	TRY	215
ASKED	398	TRAINING	211
LATER	397	NATION	209
KNEW	395	SOLDIER	206
TOWARD	386	SPEECH	206
DAYS	384	HOUR	204
PRESIDENT	382	LOT	199
SOCIAL	380	BOMB	197
NATIONAL	375	BUILD	193
FACE	371	PEACE	192
PER	371	ABLE	191
AMONG	370	READ	191
OFTEN	368	TALK	183
THINGS	368	HAPPEN	179
LOOKED	367	MILE	176
EARLY	366	PERSON	174
WHITE	365	EDUCATION	169
JOHN	362	NATURAL	167
FOUR	359	FORCE	165
WITHIN	359	MAYBE	165
FELT	357	CAUSE	163
CHILDREN	355	PLAY	162
SAW	352	LEARN	161
BEST	351	OFFICER	161
CHURCH	348	FLY	160
LEAST	343	LAND	153
DEVELOPMENT	334	PRODUCTION	153
SEEMED	332	MONTH	149
FAMILY	331	CARRY	148
MEMBERS	325	ENEMY	146
AREA	323	COLLECT	145
OTHERS	323	EVERYTHING	144
TURNED	320	SORT	143
ALTHOUGH	319	RUN	142
OPEN	319	OPERATE	141
GOD	318	ASK	140
BEGAN	312	MATCH	140
DOOR	312	BIT	139
THUS	312	REALIZE	136
SENSE	311	AIRPLANE	134
MEANS	310	BOOK	133
MATTER	308	DEVELOP	133
ITSELF	304	MAIN	132
YORK	301	AM	131
TIMES	300	FRONT	130
HUMAN	299	INDUSTRY	130
LAW	299	USUAL	130
ABOVE	296	HARD	129
ACTION	291	CANT	128
COMPANY	290	VICIOUS	127
HANDS	289	LATE	126
LOCAL	288	MACHINE	126
FIVE	286	FIRE	125
GAVE	285	DOLLAR	121
EITHER	284	PRETTY	121
ACT	283	STAND	121
ACROSS	282	ATTACK	119
THREE	282	FUTURE	117
PAST	281	STOP	117
TAKEN	281	THEREFORE	117
HAVING	279	CONTROL	117
SEEN	279	RADIO	116
DEATH	277	FINAL	116
EXPERIENCE	276	FOOD	115
WORDS	274	SUBJECT	115
ALREADY	273	DEAL	113
THEMSELVES	270	COMPLETE	112
INFORMATION	269	CONSIDER	112
SHALL	267	PICTURE	111
MONEY	265	SOON	111
PERIOD	265	WEATHER	111
HELD	264	BEGIN	110
PROBABLY	261	DEMOCRACY	110
FREE	260	SPEAK	110
REAL	260	SUPPOSE	110
SEEMS	259	BAD	109
BEHIND	258	NAVY	109
CANNOT	258	WOMAN	109
MISS	258	STUDENT	108
POLITICAL	258	YOURE	108

As substantiation of this conclusion, an inspection of the following analysis is revealing. The accompanying tables compare the occurrence frequencies of the first 400 lexical types as obtained from the vocabulary counts compiled by Kučera and Francis, Black and Ausherman and Carroll (American Heritage) for adult written, spoken and children's texts of English, respectively. Starting with the Kučera-Francis count as comparative basis, each of the remaining counts has been reordered to correspond to that count and so that any word type not within the first 400 entries of the comparision counts was deleted from the print-out. The three counts represent the most extensive and up-to-date counts of their respective types. The Kučera-Francis and Carroll tabulations are based on counts of more than one million and five million running words, respectively. The Black-Ausherman count of spoken English is based on a data base of some 288,000 total words. The three counts, thus, represent samplings of spoken, written, adult and child language and display a broad range of both stylistic and content differences. The first 400 types of each count, respectively, account for 60, 64 and 79 percent of their totals. It will be observed that with the exception of the word, YEARS, which is not within the first 400 words of the Black-Ausherman count, the first 100 entries of the Kučera-Francis count are matched by identical occurrences in the other counts. But more importantly, the order of occurrence of these matches is remarkably similar. Pearson-product moment correlations of the frequencies of these items among the three lists are all in excess of .95. In fact, even when the correlations are taken over the entire 400 word types, the correlations among the lists are still in excess of .65. Thus, despite the differences which can be expected to accrue to these differing versions of English and their respective differences in corpus size, the occurrence and order of the highest occurring types is substantially unchanged. It is also clear that these highest frequency items are nearly uniformly functors. Where differences exist in the counts, those differences with only few exceptions obtain for the substantive items. As one proceeds deeper into the frequency lists, although in absolute terms still barely into the total number of types (86,741 for the American-Heritage count), one increasingly encounters greater and greater ideational influences reflected by the differences in the data sources. For example, the Black and Ausherman count used military personnel giving extemporaneous speeches as their data base, and words such as WAR, GENERAL, SERVICE and ATTACK should not be expected to occur in a count of the language of childrens' texts for which the primary colors, numbers and body parts would be expected.

In addition to the question of optimal sample size, there still remains the question concerning the method of sampling. The samples must be sufficiently scattered with respect to subject matter so as to avoid the vocabulary biases inherent in the ideational clumping which characterizes language. Yule (1944) has specifically rejected the random strategy of sampling in favor of spread sampling. This technique spreads the sample as uniformly as possible over the whole range of the work to be sampled. Yule's suggestion was to select a sample of words from each page, the words being samples within the page unit taken either at random or from a continuous passage of a prespecified number of lines. It should be observed that the technique which we have employed for the sampling of American television in this research is a spread sample based upon randomly selected continous segments of five minute duration. The procedure is quite straightforward. A clock activates a tape recorder for a five-minute interval during each hour of total speech time. The specific five-minute interval is varied in a psuedo-random fashion so that different five-minute segments are sampled at each hour. The technique for accomplishing this sampling is instrumentally simple. The minute and hour hands of a normal clock coincide at a different locus during each hour of a day. The specific time of coincidence is given by the equation:

$$(0) \quad h{:}5h + \frac{5h}{12}$$

assuming the clock is started with the hands at 12 midnight. Thus, for example, the first coincidence of the hands would occur at 1:05.3, the second at 2:10.8 and so on. As real time progresses through the day, the five-minute sampling segments precess further into the hours. In order to avoid this consistent precession, the clock is randomly started at a different clock time each day.

Notational conventions. For convenience and consistency, the following notational definitions will be employed in this paper. A word token, i.e., any occurrence of a word unit, will be symbolized as N_w and the total number of words as simply N. Let K be the number of different words in the sample; i.e., the total types, and K_w symbolize a particular word type. For any word,

W, the symbol, W_i, shall stand for the frequency of that word in the sample. The symbol, i, as subscript indicating frequency will also be used to designate the n_i number of different words having the same i frequency. Thus, we may define the probability of a word type in a sample as the relative frequency:

$$(1)\quad p(W_i)={}^{W}i/N$$

or the probability of the n_i words of given occurrence frequency as the relative frequency

$$(2)\quad p(n_i)={}^{n}i/_{N}$$

Since the notation W_i must unambiguously refer to the frequency of a word type, (1) will alternatively be expressed as:

$$(1')\quad p(W_i)={}^{i}/_{N}$$

The fraction of the entire sample of types having frequency i will be designated as:

$$(3)\quad \theta = {}^{n}i/_{K}$$

and the fraction of the sample made up of tokens with frequency i as:

$$(4)\quad \phi = {}^{in}i/N = p(W_i)n_i$$

A frequency ordering of the W words by i, such that larger values of i have higher rank, may be ordinally transformed by assignment of ranks, r, with ties in rank given the average rank position of the equal i frequencies. Thus, the relation:

(5) $ri = C$, where C is a positive constant,

expresses the regularity first noted by Zipf of the relationship between rank and frequency of word types.

If values of n_i are plotted as a function of i one obtains a distribution known as the type-frequency distribution. Alternatively, if one plots in_i as a function of i one obtains the distribution known as the token-frequency distribution, either distribution being a word frequency distribution.

These definitions will suffice for most of the following discussion, where special symbols are introduced their definitions will be given at that time.

Four Models of Word Frequency Distribution

Four quasi-distinct models have been proposed as schematic representations of the claimed regularities of word occurrences in natural languages. These models are those proposed by Zipf, Mandelbrot and Yule and the lognormal distributions proposed by Herdan. Each has been proposed as the best schemapiric representation of the language observations and each has been critized on both schematic and empirical grounds. As we shall see, however, the lognormal model has received the most attention and, at least, at this date appears to be the more robust of the alternative formulations.

Zipf's "law". It is fair to say without severe risk that all of this began with G. K. Zipf (1935, 1949).[1] It is also not very risky to say that his contribution is probably limited to his role as originator and not to his so-called "law." Zipf's observation can be expressed simply. If each word type of a frequency sample is assigned a rank value corresponding to the decreasing frequencies of these word types and plotted as a bi-logarithmic function of those frequencies, the relationship will be roughly linear with slope of approximately -1. Zipf saw this regularity as a natural consequent of the principle of Least Effort. According to this principle, the speaker prefers "a small vocabulary that will spare him the effort involved in selecting the exact words needed to encode his message" whereas the listener prefers "a large vocabulary that will spare him the effort involved in determining which of several alternative messages the talker intended." The speaker was in Zipf's terms driven by the Force of Unification and the listener by the Force of Diversification. Zipf presumed that the equilibrium state constituting the resolution of these two opposing force, produced the rank-frequency relation. Whether or not such forces indeed exist in either the individual or aggregate language user is moot. But even if they were to exist, it is unlikely that they would supply an explanation of the particular form of the bi-logarithmic relationship between frequency and rank claimed by Zipf. The most telling criticism which by now has become almost hackneyed is that rank and frequency are of necessity lawfully related not by empirical observation but by definition. The interval scale of frequency when collapsed into the ordinal scale of rank necessarily is a negative monotone of that rank. Thus, the primary observation captured by the law is trivial. Nonetheless, it can be argued that there are an infinite number of negative monotones which are possible under the definitional relationship between ordinal and interval scales. Although the bi-logarithmic transformation proposed by Zipf defines a specific and unique such monotone, it has been demonstrated time and again, that the bi-logarithmic transform provides a very bad fit to the data at either the high or low frequency tails of the distribution. The regularity only holds for the narrow middle range of frequencies.

---

[1] In point of fact, the observation of regularity between a word's frequency and its rank in a sample had been made by both Estoup (1916) and Willis (1922) before him. Zipf, nonetheless, was largely responsible for the subsequent proliferation of theory and research on the topic.

Mandlebrot's distribution. Mandelbrot's (1953) publication of an information theory approach to language statistics constituted what amounted to a mathematically rigorous defense of Zipf's observation. Where Zipf has been vague, imprecise and mathematically naive, Mandelbrot rigorously derived the rank-frequency distribution from mathematical arguments based upon precisely defined quantities. Rapoport (1965) has provided an explication of Mandelbrot's model which because it cannot be equaled in clarity is here presented.

Consider the rank-frequency distribution, where the most frequent word has rank 1, the next most frequently occurring word has rank 2, and so on to the least frequently occurring word in the vocabulary of K words in the sample. Associated with each word is its probability of occurrence, $p(r)$, $r = 1, 2, \ldots$, where r is the rank of a specific word.

Associated with each word there is also a "cost" of producing it. The question which arises is, if the assignment of probabilities to the several

~~145~~ 146

words of vocabulary is under the control of the user of the language, which assignment will minimize the average "cost" per word? The answer is obvious - use the "cheapest" word all the time. However, by using the same word all the time no information can be conveyed from the speaker to the listener. Hence, Mandelbrot suggests using Shannon's measure of information

$$(6) \quad H = -\sum_{r=1}^{K} p(r) \log p(r),$$

where $p(r)$ are the probabilities of occurrences of the words. Formula (6) gives the amount of information per word in the sample of speech. This formula, together with that for "cost" to be given below, enables Mandelbrot to frame his law in a mathematical form.

Now, if a fixed amount of information is given per word, the question arises which frequency distribution will give minimum average "cost" per word. Or alternatively, given a fixed average "cost" per word, what will be the frequency distribution of the words which gives maximum information per word? Let $C(r)$ be the "cost" of the r-th word. The average "cost" per word is given by

$$(7) \quad C = \sum_{r=1}^{K} p(r)C(r).$$

The problem is then to maximize H subject to the constraints:

$$(i) \quad \sum_{r=1}^{K} p(r)C(r) = C;$$

$$(ii) \quad \sum_{r=1}^{K} p(r) = 1.$$

Use of LaGrange multipliers gives the equations

$$(8) \quad \frac{\partial}{\partial p(r)} \left[- \sum_r p(r) \log p(r) + \lambda_1 \sum_r p(r) C(r) + \lambda_2 \sum_r p(r) \right] = 0,$$

$$r = 1, 2, \ldots K,$$

where λ_1, λ_2 are arbitrary multipliers. The system (8) solved for each $p(r)$, give

$$(9) \quad p(r) = M^{-1+\lambda_2+\lambda_1 C(r)},$$

where M is the base to which the logarithms are taken. The constraints on the problem determine the values given to the arbitrary multipliers λ_1 and λ_2. Setting

$$(10) \quad B = M^{\lambda_2 - 1},$$

and

$$(11) \quad B = -\lambda_1,$$

it is possible to write

$$(12) \quad p(r) = PM^{-BC(r)}.$$

In order to obtain Zipf's formula from (12) it is necessary to show that $C(r)$ is a logarithmic function of the rank r. Mandelbrot does this by giving the equation for the number of words, $N(C)$, of a given cost C as:

$$(13) \quad N(C) = C(C-C_1) + N(C-C_2) + \ldots + N(C-C_G),$$

where C_1, C_2, ... C_G are the "costs" of the individual "letters." For large C the solution of (13) is approximately

$$(14) \quad C(r) \quad \log_M r,$$

where M is the largest root of the equation

$$(15) \quad \sum_{g=1}^{G} M^{-C_g} = 1,$$

and is the same M of formula (6). Substitution of (14) into (12) gives

$$(16) \quad p(r) = PM^{-B\log_M r} = Pr^{-B}.$$

This is the proposed formula for the rank-frequency distribution. Mandelbrot reports that this formula is not valid for small values of r, and suggests other formulas. A more exact formula is given by

$$(17) \quad p(r) = Q(r+m)^{-B},$$

where Q and m are constants. When $B > 1$, and $R = B$, an improvement of formula (16) is given by

$$(18) \quad p(r) = B(B-1)P^{B-1}(r+P)^{-B}.$$

Formula (18), which according to Mandelbrot "turns out to be experimentally excellent" (1961c, p. 195), was dervied by Mandelbrot from some expirical considerations. Although not presented here, Mandelbrot has also shown that the type-frequency distribution may be derived from the same initial assumptions.

The critical assumptions in Mandelbrot's model involve the notions of "cost" and "letter". By minimizing "cost", Mandelbrot establishes the relationship between it and frequency and in turn by defining that minimum in terms of "letters" as constituents of a "word" he derives the relationship between "cost" and rank. Thus, he is able ultimately to establish the desired relationship between frequency and rank through the common construct of "cost per letter." Although there is no essential requirement that these hypothetical constructs be defined in any way other than as specified by their mathematical definitions, i.e., schematically, there is considerable utility in finding psychological justification for them in an attempt to provide explanatory rather than descriptive adequacy for the assumptions. Mandelbrot has suggested several interpretations for his notion of cost, particular among these being the time required to read a word. "Letter" in turn can be either phonemes or graphemes, the total cost of a "word" becoming the sum of the constituent element costs. These elements demarked by a unique element, e.g., space, then define word.

Mandelbrot's model, sophisticated and rigorous as it is, and notwithstanding the interesting and potentially productive psychological implications implied by the notion of cost, suffers from the same criticisms which have been applied to Zipf. The empirical data still do not fit the model for extreme values of i and the parameters of the distribution are still highly correlated.

Yule's distribution. Using Yule's (1924) analysis of the distribution of the frequency of species within general classifications, Simon (1955) proposed that Yule's distribution could provide a model for the distribution of word types within frequency classifications, i.e., for type-frequency distributions. Again, following Rapoport's exposition which, in turn, closely parallels that of Simon (1955, 1957), the model can be presented as follows:

Consider a text that has reached a length of N words. First assume that the probability that the $(N+1)$-*st* word is a word that has already appeared exactly i times is proportional to in_i - that is, to the total number of occurrences of all words that have appeared exactly i times. Then assume that there is a constant probability, a, that the $(n+1)$ word is a new word - a word that has not occurred in the first N words. Given the above assumptions, Simon derives the following distribution function for the number of words used exactly i times.

$$(19) \quad n_i = n_1^* \, B(i, \frac{1}{1-a} + 1),$$

where n_1^* is the number of words which appear only once in the sample, B is the Beta function and a is a free parameter assumed to be constant for different sample sizes.

It is shown (Simon, 1957, p. 151) that

$$(20) \quad n_1^* = \frac{K}{2-a},$$

where K is the total number of types. Simon shows that when a is small, equation (19) simplifies to

$$(21) \quad n_i = \frac{K}{i(i+1)} .$$

For large i, an approximation for (21) is

$$(22) \quad n_i = \frac{K}{i^2} ,$$

which is equivalent to the Zipf relation (equation 5).

The first step in fitting the distribution expressed in equation (19) to word-count data is to get an estimate for a, the assumed constant probability that a new word will be added to the text. It would be possible, of course, to count n_1^* and to solve equation (20) and all subsequent terms may be obtained by applying a recursive formula to the Beta function:

$$(23)\quad n_i = \frac{(1-a)(i-1)}{1+(1-a)i}\ n_{i-1}.$$

There are three characteristics which are generally observed in type-frequency distributions, and which should be accounted for by any model for type-frequency distributions. First, it is observed that type-frequency distributions are J-shaped distributions with very long tails. The tails can generally be fitted by the function

$$(24)\quad n_i = \left(\frac{a}{i^m}\right) b^i ,$$

where a, b, and m are constants. Simon (1955) shows that (19) fulfills this requirement. Another characteristic of observed type-frequency distributions is that the parameter b in (24) is generally very close to 1, and m is very close to 2. In this case, (24) becomes

$$(25)\quad n_i = \frac{a}{i^2} ,$$

which is the same as (22) when a, the constant, is replaced by K. A third characteristic is that the following relations:

$$(26)\quad \frac{n_i}{K} = \frac{1}{2} ,$$

$$(27)\quad \frac{n_2}{n_1} = \frac{1}{3} ,$$

seem to hold approximately true for observed type-frequency distributions. The same relations are easily obtained from (21), ($n_1 = \frac{K}{1(1+1)} = \frac{K}{2}$, $n_2 = \frac{K}{2(2+1)} = \frac{K}{6}$, and hence $\frac{n_2}{n_1} = \frac{K \cdot 2}{6 \cdot K} = \frac{1}{3}$). The function suggested by Simon thus has all the above desired characteristics.

The model assumes a Markovian generator for which the probability of any particular word is dependent upon the probabilities of the preceeding words. A generator whose states are determined by the proceeding states has high face validity as a model of a speaker. Unfortunately, however, Simon must also assume that the state sequences are ergodic. That is to say, that the probability dependencies between states do not change over time. Simon, in fact, has confessed that "It is known empirically, at least for the most straighforward application of the model, that $\alpha (K)$, the rate at which new words appear in text, is, in fact, not a constant but a slowly decreasing function of K." (Simon and Van Wormer, 1963, p. 204). Miron and Wolfe (1964) have suggested a mechanism by which such dependencies might change over time for the lognormal model to be presented later in this paper, but it is difficult to fit such a mechanism into Yule's model.

Simon argues that psychological justification of his model can be established by assuming that language involves the processes of association and imitation. He claims that the speaker's choice of messages is determined by imitation of the messages used by other speakers and that the sequential dependencies of these messages are determined by the associations established by linguistic experience. Both processes might indifferently be applied to any stochastic model of the speaker. Both processes undoubtedly do have some influence upon what might be called speech habits, but as a general model of the speaker all such stochastic models have been shown to inadequately represent the novelty and innovation which make up the greater part of the speech process. Even were we to set aside these theoretical arguments, however, there would remain the fact that the hypothesized distribution in fact represents a rather bad fit to the empirical facts. Herdan (1962) concludes that "The discrepency between [Simon's] theory and observation is such as to invalidate Simon's claim that his model fitted word-frequency distribution in the whole range of the variable."

The lognormal distribution. Because of its particular pretinance to the theoretical justification upon which our research rests, the following earlier study by the author in collaboration with Ms. Sharon Wolfe (Miron and Wolfe, 1964) is presented here in its entirety. It provides evidence of the generality of the lognormality of the vocabulary derived from word associations elicited as reponse to stimuli embedded in linguistic frames and hence of the validity of such a procedure for ascertaining vocabulary distributions as a substitute for those procedures normally employing continuous text.

Among the several alternative theoretical laws of word-frequency distributions which have been advanced, the most recently has been the suggestion that such distributions conform to the class of skew normal distributions. Herdan (1961), notably, has brought together a series of studies the import of which is to establish that the frequency distributions of the number of words sharing the same frequency of occurrence are lognormal. Howes and Geschwind (1963) and Rapoport (1965) have used lognormal transformations of word frequencies with success in characterizing the ad libitum speech of aphasic patients and normals. And most recently, Carroll (1971) has used the Lognormal Distribution to characterize the extensive vocabulary samples complied in the production of the American-Hertigage Word-Frequency Dictionary. The present paper represents an attempt to investigate the applicability of the lognormal distribution to word-association responses in a variety of languages when those responses are restricted to qualifiers.

Kapteyn has shown that a random positive variate, the change in which is determined by a random proportion of the momentary value of the variate, will be lognormally distributed provided the assumptions necessary for the central limit theorem are met.

The probability of responses in a standard word-association task, from the viewpoint of the habit-hierarchy position, can be considered a positive variate which is the outcome of a discrete random process. Assuming, in addition, the existence of some factor or factors operating to produce momentary changes in this distribution, it is reasonable to expect that the resultant overall probability distribution should be lognormal in form. In principle, the change effect was originally postulated by Hull as influencing the momentrary response probability through the operation of an oscillatory mechanism, ${}_{S}O_{R}$. It therefore appeared appropriate to inquire whether the probability distribution of word-association responses could be shown to conform to the hypothesized distribution, with a view toward identifying the response analogues of the necessary conditions for the genesis of that distribution, if it were appropriate. In addition, it was felt that the parameters of such a distribution might reflect certain aspects of the linguistic habits of Ss from different speech communities.

METHOD

Subjects. One hundred males of high-school age in each of 12 linguistic communities were used in the study. All Ss were nominally monolingual speakers of the mother tongue of the community of which they were resident. Their ages ranged between 13 and 17. The 12 languages and places of origin of the data comprisisng the sample were as follows: Afghan-Farsi (Kabul, Afghanistan); American-English (Decatur, Illinois); Arabic (Beirut, Lebanon); Cantonese (Hong Kong); Dutch (Amsterdam, Holland); Finnish (Helsinki, Finland); Flemish (Brussels, Belguim); French (Paris, France); Iranian-Farsi (Tehran, Iran); Japanese (Tokyo, Japan); Kannada (Mysore, India); and Swedish (Uppsala, Sweden).

Testing Instrument. A standard testing procedure for eliciting qualifier associations to each of 100 stimuli was devised and appropriately modified for testing in each of the sample languages. The Ss were told to place each of the 100 substantive stimuli in a common frame sentence and to complete the frame by supplying a single qualifier which in their judgement would appropriately fit the frame. For the English-speaking Ss the frames were: "The ____________ BUTTERFLY." and "The BUTTERFLY is ___________." Both frames define Fries (1952) words of Class 3. The particular test frames or frame varied from language to language as the syntactic requirements of qualifier distribution varied.

The 100 substantives were selected from a pretested pool of 200 items. The original 200 items in turn were drawn in part from a list of items used in glottochronological investigations purported to be of wide linguistic applicability from the Kent-Rosanoff list and from category headings used by the Human Resources Area Files index.

Testing Procedure. Testing was carried out in the class rooms of the schools, Ss being run in large groups comprising the normal class. Complete testing for the 100-stimuli final list required approximately one-half hour. The Ss were instructed to attempt to supply an associate to all items but to omit any items with which they experienced inordinate difficulties. The 100-item final stimulus list was derived from the tests administered in Finnish and English for the 200-item lists, the procedures being identical to those used in later tests except for the increased length of task in these two countries.

Scoring Criteria. All associated responses were inspected by native speakers of each of the languages for non-conformity to instructions. Responses judged not to be admissible within the test frames were discarded. Such discarding of responses, however, was kept to an absolutely small limit, doubtful instances being retained. All grammatical inflections and orthographic variants were regularized to a single consistent form, with only difference in root forms being considered instances of separately distinct responses. Combinations of free morphemes, multiple-word or phrase responses and neologisms were accepted. In all instances the assumption was made that S's response was acceptable unless the response in question was clearly deviant from minimal usage standards.

Method of Analysis. Each qualifier type has an associated frequency of occurrence representing the total number of occurrences of the type across all Ss over all stimuli. The types can be classified by occurrence frequency: category i contains all n_i types which share occurrence frequency f_i. It was hypothesized that the distribution of the random variable F, which takes on values f_i, is lognormal; i.e., if the variable $X = \log F$ is introduced, the distribution of X is normal. This is expressed in the equation $P(X \phi \log f_i) = \phi (z_{\log f_i})$, where ϕ is the standard normal cumulative distribution, f_i is a particular occurrence frequency, and $z_{\log f_i} = (\log f_i - \mu/\sigma)$. Thus, the probability of obtaining a category with frequency of occurrence 500, for instance, is given by $\phi (z_{\log 500}) - \phi (z_{\log 499})$. Another way of interpreting this statement is to say that the probability associated with occurrence frequency 500 is simply the proportion of types which are expected to have occurrence frequency 500. The empirical estimate of this probability, in turn, is simply the number of types actually occurring with frequency 500 divided by the total number of types.

These empirical estimates of the probabilities were used to obtain least-squares estimates for μ and σ, the parameters of the normal distribution. For category i, $\Sigma^{i}_{j=1} n_j/n_t$ gives an estimate of the cumulative probability, which can then be transformed with the aid of standard normal tables to a z-score. This transformation in turn yields a set of empirical z-scores, where $z_i = (\log f_i - \mu)/\sigma$. This equation is linear in both of the variables, Z and $X = \log F$; accordingly the least-squares solutions to the general linear equation $Z = mX + k$ were obtained such that $m = 1/\sigma$ and $k = -\mu/\sigma$. Once the least squares solutions μ and σ were obtained, predicted z-scores, $Z = (X - \mu)/\sigma$ were calculated, and the predicted cumulative probabilities $\phi(Z)$ were obtained from normal tables.

It can be shown that if the distribution of occurrence frequencies for types is lognormal, the first moment of this distribution defines the distribution of occurrence frequencies for tokens. The total number of tokens in category i is simply the product of f_i, the number of tokens for each type, and n_i, the number of types sharing this occurrence frequency. Here again, $P(X < \log f_i) = \phi(z_{\log f_i})$. However the estimate of ϕ here is given by $\Sigma^{i}_{j=1} n_j f_j$, the total number of tokens occurring in categories of occurrence frequency f_i or less, divided by the total number of tokens, $\Sigma n_j f_j$.

BEST COPY AVAILABLE

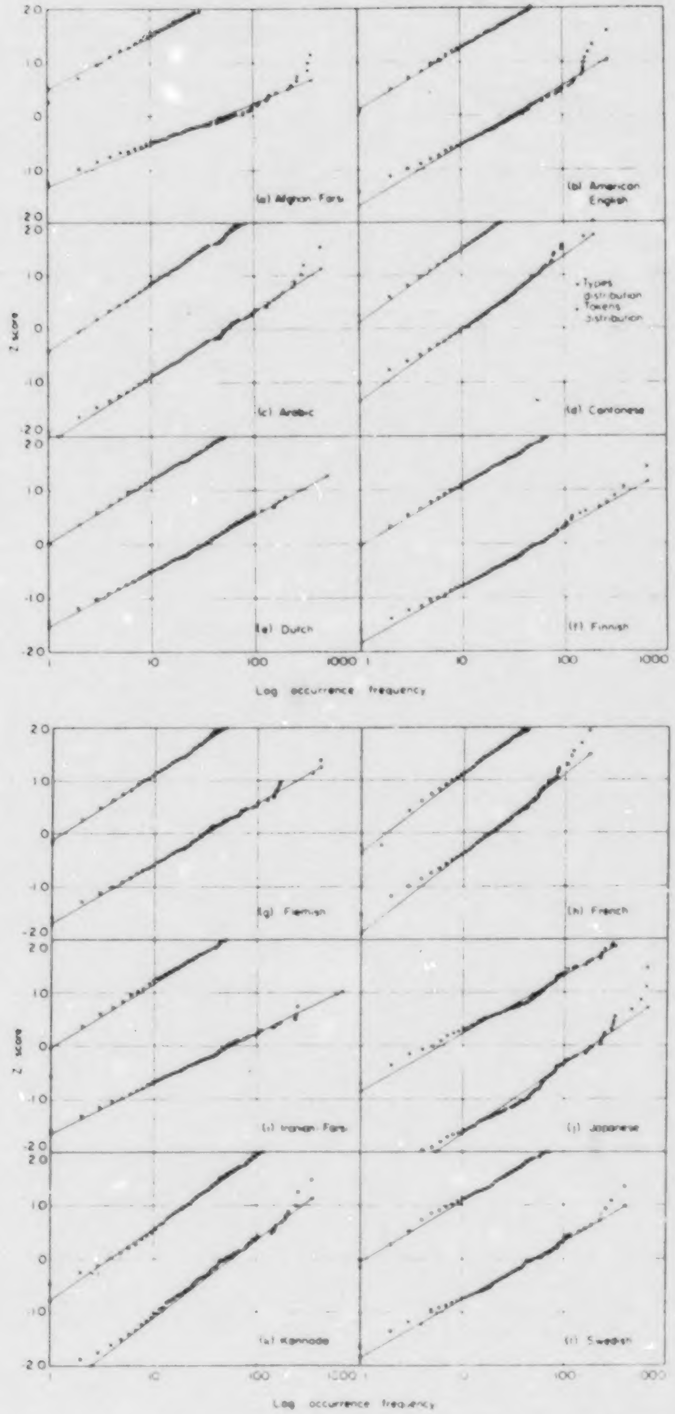


FIG. 1, (a-l). Distributions of qualifier types (upper line) and tokens (lower line) for the twelve language samples.

BEST COPY AVAILABLE

The least-squares estimates for μ and σ for the token distributions were calculated by the same method as for types.

As a further test of the lognormality of the distributions for types and for tokens, it can be shown (see Herdan, 1961) that the *j*th moment of a lognormal variate with parameters μ_{oe} and σ^2 is also a lognormal variate with parameters $\mu_j = \mu_o + j\cdot\sigma^2$ where logarithms are taken with respect to the base *e*. For this study, with $X = \log F$, the mean of the *j*th moment becomes $\mu_j = \mu_o + \log_e 10\cdot j\cdot\sigma^2$. Thus the variance for tokens should equal the variance for the types distribution, and the means should show the relationship expressed above; since $j = 1$ in this case, the equation becomes $\mu_{tokens} = \mu_{types} + \log_e 10\cdot\sigma^2_{types}$.

RESULTS

Figure 1 displays the primary results of the foregoing analysis for each of the language samples. If lognormality holds and if the f_i are plotted against cumulative proportions on lognormal paper, a straight-line graph should result. Although none of the extant significance tests commonly employed for estimating goodness of fit is entirely appropriate for functions of this kind, inspection of the figures clearly indicates sensible linearity for a major proportion of the transformed empirical points. Correlation co-efficients computed between predicted and obtained *z*-scores ranged between 0.900 and 0.999 for the type distributions and between 0.900 and 0.998 for the token distributions. Since the squared correlation coefficient gives

BEST COPY AVAILABLE

TABLE 1

ESTIMATED LOGNORMAL PARAMETERS, TRANSFORMATION CONSTANTS, AND NUMBER OF DETERMINATIONS ON WHICH BASED

Language	Types							Tokens					
	N^a	n	m	k	μ	σ	μ/σ	N	m	k	μ	σ	μ/σ
Arabic	708	71	1.23	—0.40	0.33	0.81	0.41	8749	1.22	—2.12	1.74	0.82	2.12
Cantonese	2509	65	1.38	0.10	—0.07	0.73	—0.10	9730	1.35	—1.36	1.01	0.74	1.36
Dutch	1442	68	1.20	0.04	—0.03	0.83	—0.04	9557	1.04	—1.55	1.49	0.96	1.55
English	1138	62	1.12	0.13	—0.12	0.90	—0.13	7349	1.11	—1.66	1.49	0.90	1.66
A-Farsi	1681	36	1.00	0.50	—0.50	1.00	—0.50	9582	0.79	—1.33	1.68	1.26	1.33
I-Farsi	1117	43	1.19	—0.04	0.03	0.84	0.04	9612	0.93	—1.65	1.77	1.07	1.65
Finnish	1072	73	1.12	—0.03	0.03	0.90	0.03	9768	1.08	—1.87	1.73	0.93	1.86
Flemish	1260	66	1.25	—0.13	0.10	0.80	0.13	9572	1.13	—1.69	1.50	0.89	1.69
French	1566	67	1.45	—0.34	0.23	0.69	0.33	9813	1.50	—1.86	1.24	0.67	1.85
Japanese	290	81	1.13	—0.85	0.75	0.89	0.84	9997	1.33	—2.97	2.22	0.75	2.96
Kannada	612	75	1.34	—0.78	0.58	0.74	0.78	9600	1.39	—2.42	1.74	0.72	2.42
Swedish	903	68	1.13	—0.04	0.03	0.88	0.03	7881	1.09	—1.84	1.68	0.92	1.83

[a] N = Number of word types or tokens obtained, n = number of occurrence categories, and m and k are constants in the linear equation $Z = mX + k$.

the proportion of variance of the obtained distribution accounted for by the predicted distribution, the fit would seem to be remarkably good. It is still entirely possible, however, to find that the remaining variance, despite its small size, is large relative to the error and hence significant. At least three considerations, however, militate against attempting to test the significance of the departure of the data from the hypothesized distributions. First, the values of the occurrence frequency variable cannot be considered independent. Second, there are pronounced end-effect distortions in these distributions due to the finite size of the sample and to the finite and variable step increments of the occurrence frequencies. Also, the precision of estimates of the probabilities for each occurrence frequency is necessarily greater for categories containing large numbers of responses than for those categories containing few responses; accordingly the various occurrence-frequency categories should not be given equal weight in estimating departures from lognormaility. The third argument involves the logic of significance testing. Given the impressively large proportion of variance accounted for by the hypothesized distribution, it does not seem reasonable to test the unexplained variance. For practical purposes, the best estimate of the position of an undetermined point would be that provided by the parameters of the fitted curves, the least-squares estimates of which are displayed in Table 1.

It will be observed that the language distributions displaying greatest differences in slopes (variances) are those for Afghan Farsi and French. Of all the languages, Afghan Farsi exhibited the flattest slope (highest variance) for both type and token distributions, indicating for this group of respondents that responses tended to be evenly distributed across occurrence categories. The French data, on the other hand, tended to have greater variation in numbers of responses in the different frequency-of-occurrence categories; i.e., the French distributions displayed lowest variance and steepest slope. Inspection of the μ estimates indicates that the difference in variances of the distributions for these two languages is apparently attributable, in part, to the greater number of single-occurrence responses in the French data.

The preferential emmission of either low- or high-frequency qualifiers undoubtedly has its basis in either or both linguistic and cultural characteristics of the speech community. Samples with high mean occurrence frequencies reflect a preferential usage of qualifiers elicited from most of the other respondents and stimuli, while samples with low means reflect predominate usage of qualifiers idiosyncratic to the individual or to particular stimuli. Given equal variances, therefore, those samples with highest mean occurrence frequencies can be characterized as exhibiting greater stereotypy of response than those with lower means.

Since stereotypy seems to be a function of both parameters, a working definition for this concept is given by the ratio μ over σ as in Table I. These ratios indicate that the type distributions for Afghan Farsi, English

BEST COPY AVAILABLE

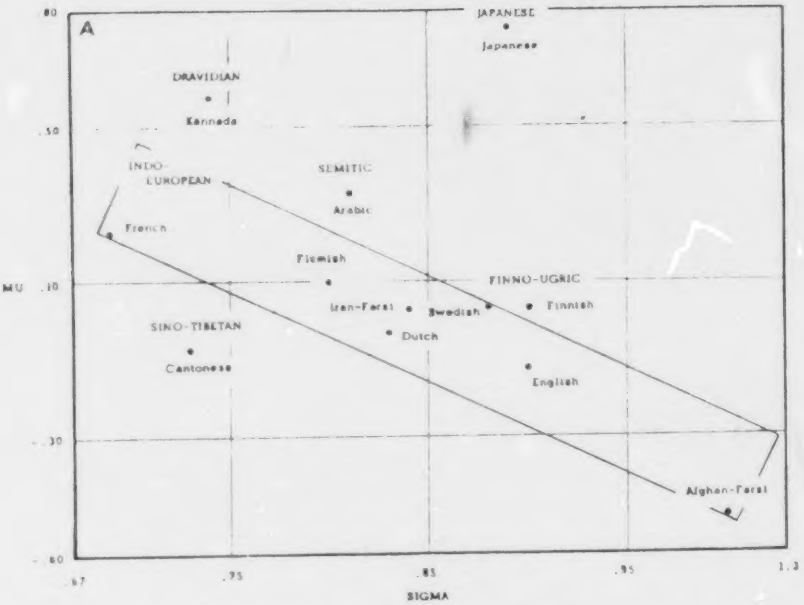


FIG. 2 (A and B). Co-variation of the estimated lognormal parameters of the qualifier type distributions. Note: Language families in caps.

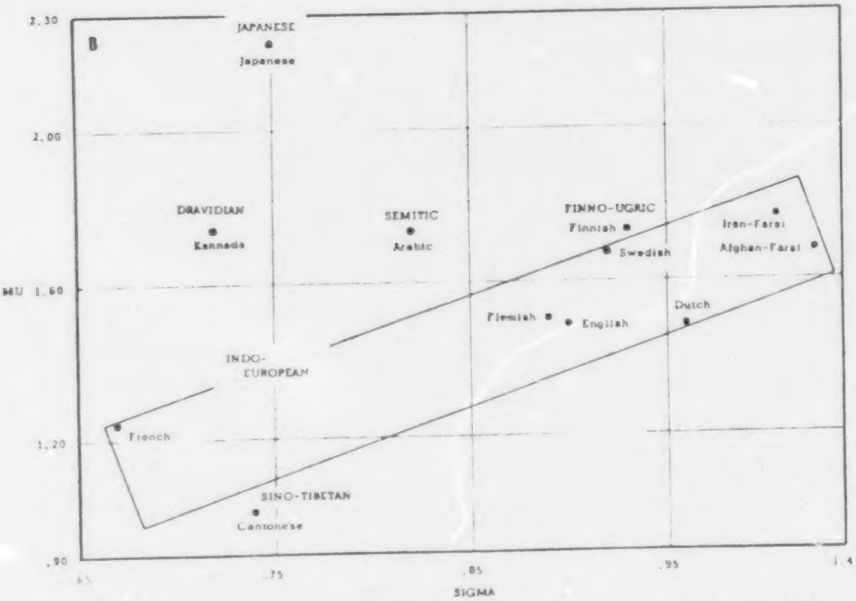


FIG. 3. Co-variation of the estimated lognormal parameters of the qualifier type (A) and token (B) distributions. Note: Language families in caps.

Cantonese, and Dutch respondents have lowest stereotypy in that order, and the Japanese, Kannada, Arabic, and French respondents displayed greatest stereotypy in that order. For the token distributions the respondents with least stereotypy in decreasing order were: Afghan Farsi, Cantonese, Dutch, and Iranian Farsi. Those with greatest stereotypy in decreasing order were: Japanese, Kannada, Arabic, and Finnish.

Hartley's test of the differences in variances and an analysis of variance applied to the differences in means between language samples indicated that both parameters of the individual lognormal distributions differentiated among the languages when considered in the aggregate. The joint parameter variations for the type and token distributions are displayed in Figures 2 and 3, respectively.

Since the token distribution should represent the first moment of the distribution for types, the two lines for a given language should be parellel (i.e., have equal variances), and separated by a distance of $\log_e 10\cdot\sigma^2$. The μ value for tokens and that computed from the relationship of the moments of the lognormal distribution specified by the equality $\mu_1=\mu_o+\log_e 10\cdot\sigma^2$ ranged betweem 0.76 and 0.05. The value for σ^2 which was used in these computations was obtained by averaging the variances of the type and token distributions for each language. Due to the absolutely small magnitude of the standard error, all of these differences are significant beyond the 1% level, both for the individual languages as well as for all languages in aggregate, when tested by t. Thus, although the magnitude of the separation between the means closely approximates that predicted by theory, significance tests indicate that these departures from predicted values are significant.

DISCUSSION

Previous successful applications of the lognormal transformation to word-frequency distributions have been made on nouns, function words, and all word occurrences in running texts (see Herdan, 1961). Apparently we may now add qualifiers as obtained from a restricted word association procedure to the growing list of such successful applications. In view of the comparability of occurrence-frequency estimates obtained in word-association procedures and other methods of obtaining word counts it would have been surprising to find lack of fit in this study where other investigators have demonstrated lognormality of the frequency of various word classes obtained from running texts.

No compelling linguistic explanation for the obtained ordering of the groups in terms of their joint parameter variations appears to exist. Although three of the low-stereotype languages are of Indo-European origin, the presence of a Sino-Tibetan language (Cantonese) in this group makes it unlikely that language-family differences alone will suffice. On the other hand, there is a clear orderly arrangement of the joint parameter variations for the languages of the Indo-European family. The values of μ and σ for the Indo-European languages are reasonably well fitted by a linear expression. In order to explain the variations within language family, however, it is necessary to invoke some additional explanation. It is possible that these variations might be explained by the nature of student-teacher interactions. Variation in the formality of secondary school education would be expected to influence the amount of innovation the students perceive as permitted. It is of interest to note in this regard that the Japanese sample, for which a high degree of formality in secondary education is a well-attested fact, exhibits the highest stereotypy values.

The stereotypy groupings presumably must involve a difference in the relative habit strengths of the responses to the stimulus aggregate. For the low-stereotypy group, this would indicate that more nearly equal habit strengths are present in the various S-R habit hierarchies than exist in the high-stereotypy subject-stimulus group as a whole. The high-stereotypy group would be characterized by one or more highly dominant responses, with the remaining responses of the divergent set being of disparately low probability. Such differences in the structure of S-R hierarchies can be attributed to differences in the predominance of certain linguistic conventions. If this is the case, we should expect to find that the high-stereotypy languages should exhibit more conventionalized conformity in sequences of modifier-substantive usage in the language as a whole.

If the response-frequency variate of this study can be considered an index of the probabilities of responses in the response hierarchy of the aggregate subject-stimuli combinations, and if these probabilities are subject to random change, it should be possible to specify the nature of the change operation in such a way as to derive the lognormality of the variate. Let us hypothesize that the stimulus aggregate produces an ordered set of responses varying in probability of emission. Let us further specify that from moment to moment these probabilities are subject to change, that change being attributable to an hypothesized tendency on the part of S to condition his choice of responses on the basis of the responses given to the previous items on the list.

Thus, although a given response may originally reside in a category of high probability of occurrence to the aggregate stimulus set, its pre-occurrence to one of the items of the stimulus set places it in a variably lower probability category. Subsequent repeated usage of the same response, in turn, places the response in increasingly lower probability categories, the extent of the probability decrement being variable across both individuals and responses. The proportion of reduction in probability of the re-emission of a specific response in a task of this kind is considered to be a constant fixed for the individual by the linguistic habits of his community, the size of his vocabulary, and his individual sensitivity to repetition within the limits established by the community. Although the proportional probability decrements for responses are constant for any given individual, their distribution across Ss is assumed to be randomly proportional to the probability of the last value of the occurrence-frequency variate. Accordingly, the parameter variations of this study provide an index of the average value of these individual-subject variations for the languages investigated.

SUMMARY

One-hundred Ss in each of 12 widely divergent linguistic communities were administered a standarized restricted word-association test consisting of 100 substantive stimuli. The Ss were instructed to provide a single response which conformed to the requirements of substitutability in a test frame designed to restrict responses to qualifiers only. The total frequency of all unique responses, excluding grammatically inflected responses, was tabulated. Categories of equal frequency of occurrence were determined and the distribution of the number of responses sharing the same frequency of occurrence was plotted. It was hypothesized that these distributions should substantially conform to a theoretical distribution of the lognormal form, since many aspects of the word-association task have high similarity to the generative rules of the lognormal distribution.

The obtained distributions were found to conform sensibly to the hypothesized distribution. An analysis of the variance explained by the lognormal equations of best fit to the transformed points indicated that very little variance remained unaccounted for by the hypothesized distribution. Accordingly, variations of the estimated parameters were examined for clues as to the nature of the processes these parameters might reflect. The concept of stereotypy of response was introduced and defined as the degree of response uniformity across both Ss and stimuli.

More generally, it was suggested that this stereotypy could be expected to be the result of stable linguistic conventions. The individual's responsiveness to these conventions was assumed to be a function of his sensitivity to response repetition within the limits established by the speech community.

The American-Heritage Project. The most ambitious test of the lognormality of word frequency distributions is that recently completed with the publication of the American-Heritage *Word Frequency Book* (1971). Under the direction of John Carroll, more than 5 million total words were sampled from approximately 1000 graded school texts and reading sources. This corpus, called the American-Heritage Intermediate Corpus (AHI), formed the source material for a successful test of the lognormal model. The entire project has resulted in what must be adjudged one of the most useful and elabrorate of statistical analyses ever completed. Carroll states: "As one inspects the data assembled here, many questions come to mind: How representative of the total lexicon of English are the word types that are listed? How accurate and reliable are the frequency data? How do the vocabularies for the various grade levels and subject matters differ? What is the effect of the word-unit chosen to be the basis of the frequency counts?

To some of these questions it is now possible to give answers that are probably correct within fairly narrow limits. Many of these answers can be derived through the analysis of the Corpus on the basis of a powerful statistical model of vocabulary that can be shown to account for the data in a surprisingly prescise way. This model, which apparently was first developed by G. Herdan, is called the *lognormal* model, because it postulates that the total vocabulary underlying a corpus is distributed according to the familiar 'normal distribution' when the logarithms of the frequencies are used."

Having accepted the lognormality of the AHI corpus, Carroll is able to predict the probabilities of the word token and type occurrences in the assumed total population of the English language. In addition, one can determine the expected number of word tokens which will be accounted for by any given number of word types and the relative frequency of occurrence of each. All of this is, of course, dependent upon the assumption that the lognormal model is, in fact, an adequate schemapiric representation of the data and that the data itself is an adequate representation of the English language. The first assumption has received more empirical support than the alternative models which have been proposed for word frequency distributions. As this paper has tried to indicate, however, such support can never constitute a proof. A schemapiric model of any domain can only be judged in terms of the theoretical desiderata of parsimony, productiveness, explanatory adequacy and utility. What Carroll, Miron and Wolfe and others have acheived is a demonstration of the descriptive adequacy of such a model as an account of their data observations. The results of the remaining tests of that model's adequacy still must be considered to be only tentatively suggestive.

The second assumption clearly has some limitations in the Carroll test, as indeed, it does in all of the other tests, although in differing ways. The AHI corpus is based on a sampling of texts typically employed by third through ninth graders. This corpus produces a theoretical expectation that the English language contains a total of 606,906 word types as estimated from the empirical occurrences of the 86,741 types actually appearing in the AHI sample. If this population estimate is to have any use, it necessarily implies that (1) new words which enter the language must displace old words, (2) that the population growth rate of English is fixed; i.e., that the birth rate exactly equals the death rate of words and (3) that a lexicon of this size will exactly specify the "character" of English within any specified and arbitrarily small value of precision approaching zero as a limit. At the least, one would want to hedge one's faith in these implications by the caveats that it is (1) the child's English which is being addressed, (2) that "type" has been defined by orthographic pattern (e.g., *word*, *words*, *wording*, *Word* are distinct type entries in the AHI) and that (3) the writer's of books for schools undoubtedly have already assumed a limitation on the vocabularies of their readers. Nonetheless, the AHI analysis represents the closest approximation yet achieved to a precise specification of the vocabulary characteristics of English, the caveats notwithstanding.

Considering the expected users of the AHI data analyses, the exposition of the procedures is extraordinarily complex. But, if the exposition

BEST COPY AVAILABLE

Table D-2. An illustration of the computation of *F*, *D*, *U*, and SFI for combined types.*

Subject Category		Type							Total (f_i)	Total Tokens in Category (s_i)	$s_i/N = t_i$	$f_i/s_i = p_i$	$-p_i \log p_i$
		word	*Word*	*word's*	*worded*	*wording*	*words*	*Words*					
Read	A	1518	3	0	0	2	1200	3	2726	**1182971**	**.232467**	.002304	.006078
Eng & Gr	B	1281	0	2	1	2	1930	3	3219	**283367**	**.055685**	.011360	022091
Comp	C	105	0	0	0	1	183	0	289	**57776**	**.011354**	.005002	011509
Lit	D	120	2	0	0	0	192	0	314	**277907**	**.054612**	.001130	003330
Math	E	120	0	0	0	0	154	0	274	**387619**	**.076171**	000707	.002227
Soc Stud	F	96	0	0	1	0	93	0	190	**503620**	**.098967**	000377	001292
Spell	G	3782	266	0	0	0	6680	485	11213	**210157**	**.041298**	053355	.067912
Sci	H	101	0	0	0	0	69	0	170	**510570**	**.100333**	.000333	.001158
Music	J	82	0	0	0	0	324	0	406	**209364**	**.041142**	.001939	.005260
Art	K	12	0	0	0	0	35	1	48	**47887**	**.009410**	.001002	.003006
Home Ec	L	7	0	0	0	0	6	0	13	**83387**	**.016386**	.000156	.000594
Shop	M	3	0	0	0	0	12	0	15	**65375**	**.012847**	.000229	.000835
Lib F	N	65	0	0	0	2	96	0	163	**303603**	**.059661**	.000537	.001756
Lib NF	P	103	0	0	0	0	112	0	215	**374885**	**.073669**	.000574	.001859
Lib Ref	Q	63	0	0	0	0	57	0	120	**271040**	**.053262**	.000443	001485
Mag	R	69	0	0	0	0	67	0	136	**314643**	**.061831**	.000432	001454
Relig	S	5	1	0	0	0	5	0	11	**4595**	**.000903**	.002394	.006274
Total		7532	272	2	2	7	11215	492	19522 $= F$	**5088766** $= N$	$\Sigma f_i t_i =$ 1411.13 $= f_{min}$	.082274 $= P$	.138120 $= -\Sigma p_i \log p_i$
D		.5197	.1614	.0000	.2306	.4037	.4756	.0312	.4828				
U		829.53	10.57	.0219	.1140	.6382	1124.2	6.993	1995.42				
SFI		69.2	50.2	23.4	30.6	38.0	70.5	48.5	73.0				

$$\log n = \log 17 = \mathbf{1.230449} \qquad \log P = -1.084737$$

$$D = [\log P + (-\Sigma p_i \log p_i)/P]/\log n = [-1.084737 + .138120/.082274]/\mathbf{1.230449} = .4828$$

$$U = (1000000/N)[FD + (1 - D)f_{min}] = \mathbf{.196511}\,[(19522)(.4828) + (1 - .4828)(1411.13)] = 1995.42$$

$$\text{SFI} = 10(\log U + 4) = 10(3.300034 + 4) = 73.0003$$

---

*In this table, certain conventions are followed. A single-line rule within the table indicates that a sum of a column or row of figures is taken. Numbers printed in boldface are numbers that may be considered as constants that would always be used in computations involving combinations of frequencies over the 17 subject categories of the AHI Corpus; they do not need to be recomputed unless one desires to combine frequencies over fewer than 17 categories. However, the formulas are given in general terms to permit their use in any desired computations.

The type frequencies for the seven types are extracted from the Alphabetical List; in addition, the several values of *D*, *U*, and SFI for these types as given in the Alphabetical List are shown at the bottom left for comparison with the values computed here for the seven types combined.

The total frequencies for the seven types are computed by simple addition across the rows and listed in the column headed "Total (f_i)." The next column gives the total number of tokens in each subject category (s_i). From this point on, the calculations are self-explanatory. The final results are shown at the bottom of the column headed "Total (f_i)."

It may be noted that the value of *U* for the combined types, 1995.42, is approximately equal to the sum of the separate *U* values, 1972.06. The amount of the discrepancy is a function of the extent to which the frequencies are differently distributed over the subject categories in the several types. Thus, in general it is legitimate to sum *U* values for separate types if it is kept in mind that the result will be only an approximation.

is difficult to follow, the editors can find excuse in the difficulty of their task. Consider the following expository table (American-Heritage Word Frequency Book, TABLE D-2, page 3) illustrating the calculations of F, the frequency of occurrence; D, the diversity or dispersion of occurrence; U, the estimated frequency of occurrence per million tokens adjusted for diversity of occurrence; and *SFI*, the standard frequency index of type occurrence per token based upon these prior calculations.

SFI corrects the occurrence probabilities for dispersion across the differing content samples which make up the AHI corpus. This correction employs the information theory measure of uncertainty and weights more heavily those word types which are more nearly equal in frequency across differing content samples. *SFI*, in turn is related to occurrence probability by the relation:

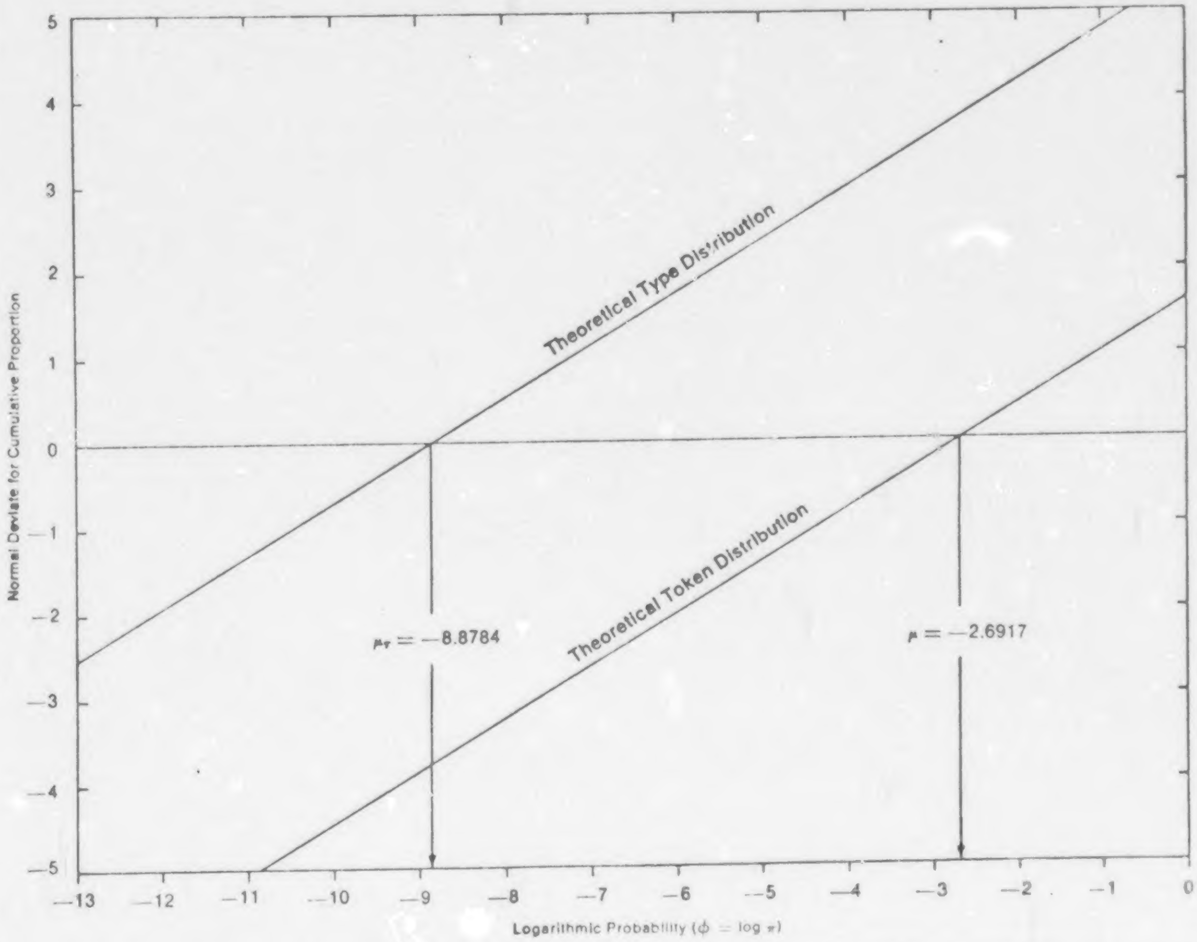
$$(28) \quad SFI = 10[(\log P)+10]$$

For our purposes, it is the lognormality of the probabilities of the token and type occurrences which will be considered. Carroll's Figure B-3 (American-Heritage Word Frequency Book, page xxv) graphically represents the theoretical cumulative word-frequency distributions which best model this data. The cumulative proportion of the total type and token distributions are plotted on the ordinate as normal deviates and the log probability of occurrence along the abscissa.

BEST COPY AVAILABLE

Figure B-3. Representation of the theoretical type and token distributions of Figure B-2 on lognormal coordinates. The proportions of areas below and above selected values of normal deviates are as follows:

Proportion Below	Normal Deviate	Proportion Above
.9999997133	5	.0000002867
.99996833	4	.00003167
.998650	3	.001350
.97725	2	.02275
.8413	1	.1587
.5000	0	.5000
.1587	—1	.8413
.02275	—2	.97725
.001350	—3	.998650
.00003167	—4	.99996833
.0000002867	—5	.9999997133



It is from this model that one can asses the relationship of number of occurrence tokens to number of occurrence types. Assume that we wish to estimate the theoretical number of types which would be required to account for 50% of all of the token word occurrences in English. Entering the ordinate of Figure B-3 at the value of zero which corresponds indifferently to the mean, median and mode of the symmetrical normal distribution, we find that this cumulative proportion corresponds to a theoretical token probability of .04917 (antilog of -2.6917 = .04917). That is to say, we should expect to find 50% of all words of English occurring with frequencies of 4 or greater per 100 words.

If we now enter the abscissa at a value which corresponds to this probability and find the corresponding point on the ordinate which intersects the type distribution at that abscissa value, we can ascertain the cumulative proportion of types which would have probabilities of 4 or greater occurrences per 100 words. Or stated otherwise, the ordinate value of the type distribution corresponding to the 50% point of the cumulative token distribution represents the number of types theoretically represented in 50% of all token occurrences. The cumulative normal value in this instance is approximately 3.7 as estimated from the graph. A normal deviate of this magnitude means that approximately .0005 times 100 percent of all types have occurrence probabilities equal to or greater than the value of .04917. Assuming the theoretical total number of types in Englsih to be 606,906, we determine that approximately 300 word types should theoretically account for 50% of all word occurrences (.0005 times 606,906).

It one were to attempt to account for 95% of all word occurrences, the same procedure would result in an estimate that some 10,000 word types would be required. And for 99%, 44,000 types would be required. In each instance, it is the highest frequency types which must be selected and the best estimate of which particular types these may be is derived from their ranks in the sample as determined by their respective *SFI* values.

Concluding Observations

As we have indicated, all of the foregoing is as weak or as strong as are the assumptions upon which the models rest. It is our belief and the substance of our recommendation that the elicitation procedures which we have outlined and which form the basis of our research have strong justification for their assumptions. Further, it would seem that those procedures largely obviate the conceptual abstractions of the data which Carroll, for example, is required to make in order to satisfy the assumptions of the model he employs. We require only the assumption that the speaker of a language will "spew" his vocabulary in an order which is isomorphic with the probabilities of those vocabulary items in the language he speaks. In order to select those items which have greatest utility over as wide a linguistic context as possible, we conceptually abstract the notion of high information over stimulus environments within a form class frame. This is the equivalent of the abstraction which Carroll's *SFI* makes with respect to content sources. It differs in that in our research we have defined informational uncertainty in terms of differences across the speakers of a defined subject population rather than across the authors of differing content texts.

SELECTED BIBLIOGRAPHY

Aborn, M., Rubenstein, H., and Sterling, T.D. Sources of contextual restraint upon words in sentences. *J. exp. Psychol.*, 1959, *57*, 171-180.

Aitchison, J. and Brown, J.A.C. *The Lognormal distribution*. London: Cambridge Univer. Press, 1957.

Bartlett, M.S. The statistical significance of canonical correlations. *Biometrika*, 1941, *32*, 29-38.

Bloomfield, L. *Language*. New York: Holt, 1933.

Bock, R.D. and Jones, L.V. The measurement and prediction of judgmental response. Chapel Hill: The Psychometric Laboratory, 1963, unpublished manuscript.

Bosch, A.J. The Polya distribution. *Statistica Neerlandica*, 1963, *17*, 201-213.

Brass, W. Simplified methods of fitting the truncated negative binomial distribution. *Biometrika*, 1958, *45*, 59-68.

Bush, R.R. Estimation and evaluation. In R.D. Luce, R.R. Bush, and E. Galanter (Eds.) *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1963, 429-491.

Bush, R.R. and Estes, W.K. (Eds.) *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959.

Bush, R.R. and Mosteller, F. A comparison of eight models. In R.R. Bush and W.K. Estes (Eds.) *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959, 293-307.

Cassirer, E. *Philosophie der symbolischen formen*. Berlin: B. Cassirer, 1923-1929.

Cherry, C. *On human communication*. New York: Wiley, Science Editions, 1961.

Chomsky, N. *Syntactic structures*. The Hague, Mouton, 1957.

Chomsky, N. Belevitch, V. Langage des machines et langage humain: a review. *Language*, 1958, *34*, 99-105.

Chomsky, N. J.H. Greenberg: essays in linguistics: a review. *Word*, 1959, *15*, 202-218.

Cofer, C.N. (Ed.) *Verbal learning and verbal behavior*. New York: McGraw Hill, 1961.

Cofer, C.N. and Musgrave, B.S. (Eds.) *Verbal behavior and learning*. New York: McGraw Hill, 1963.

Eldridge, R.C. *Six thousand common english words*. New York: privately printed at Niagara Falls, 1911.

Estoup, J.B. *Gammes stenographiques*. Paris: Institut Stenographique de France, 1916.

Feller, W. *An introduction to probability theory and its applications*. Vol. 1, (2nd ed.), New York: Wiley, 1957.

Fillenbaum, S. and Jones, L.V. An application of "cloze" technique to the study of aphasic speech. *J. abnor. soc. Psychol.*, 1962, *65*, 183-189.

Fillenbaum, S., Jones, L.V., and Rapoport, A. The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. *J. verbal learn. verbal behav.*, 1963, *2*, 186-194.

Fillenbaum, S., Jones, L.V., and Wepman, J.M. Some linguistic features of speech from aphasic patients. *Language and Speech*, 1961, *4*, 91-108.

Fisz, M. *Probability theory and mathematical statistics*. New York: Wiley, 1963.

Fries, C.C. *The structure of English*. New York: Harcourt, Brace, 1952.

Gnedenko, B.V. and Kolmogorov, A.N. *Limit distributions for sums of independent random variables*. (Translated by K.L. Chung and J.L. Doob.) Cambridge, Mass: Addison-Wesley, 1954.

Goodman, L.A. Kolmogorov-Smirnov tests for psychological research. Psychol. Bull., 1954, *51*, 160-168.

Greenberg, J.H. *Essays in linguistics*. Chicago: The Univer. of Chicago Press, 1957.

Greenwood, M. and Yule, G.U. An inquiry into the nature of frequency distributions representative of multiple happenings, etc. J. Roy. Stat. Soc., 1920, 83, 255.

Hald, A. Statistical theory with engineering applications. New York: Wiley, 1954.

Herdan, G. Language as choice and behavior. Groningen: Noordhoff, N.V., 1956.

Herdan, G. Type-token mathematics. The Hague, Mouton, 1960.

Herdan, G. A critical examination of Simon's model of certain distribution functions in linguistics. Applied Statistics, 1961, x, No. 2.

Herdan, G. The calculus of linguistic observations. The Hague, Mouton, 1962.

Herdan, G. Quantitative linguistics or generative grammar? Linguistics, 1964, 4, 56-65.

Horvath, W.J. A stochastic model for word association tests. Psychol. Rev., 1963, 70, 361-364.

Horvath, W.J. and Foster, C.C. Stochastic models of war alliances. J. Conflict Res., 1963, 7, 110-116.

Howes, D. and Geschwind, N. Statistical properties of aphasic language. Report of the Psychology Group of M.I.T., 1962.

Ijiri, Y. and Simon, H.A. Business firm growth and size. Carnegie Tech., 1963, unpublished manuscript.

Jaffe, J., Cassotta, L., and Feldstein, S. Markovian models of time patterns of speech. Science, 1964, 144, 994-886.

Jones, L.V., Goodman, M.F., and Wepman, J.M. The classification of parts of speech for the characterization of aphasia. Language and Speech, 1963, 6, 94-107.

Johnson, N.L. Systems of frequency curves generated by methods of translation. Biometrika, 1949, 36, 149-176.

Kalecki, M. On the Gibrat distribution. Econometrica, 1945, 13, 161-170.

Kapteyn, J.C. *Skew frequency curves in biology and statistics.* Astronomical Lab., Groningen: Noordhoff, 1903.

Lees, R.B. Apostel, L., Mandelbrot, B., and Morf, A. Logique, langage et theorie de l'information: **a review.** *Language*, 1959, *35*, 271-303.

Levine, L. Essentials of linguistic analysis. Chapel Hill: Univer. of North Carolina, 1964. Unpublished manuscript.

Luce, R.D., Push, R.R., and Galanter, E. (Eds.) *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1963.

Mandelbrot, B. An informational theory of statistical structure of language. In W. Jackson (Ed.) *Communication theory*. London: Butterworths, 1953.

Mandelbrot, B. On the language of taxonomy: an outline of a 'thermostatical' theory of systems of categories with Willis (natural) structure. In C. Cherry (Ed.) *Information theory*. London: Butterworths, 1956, 135-145.

Mandelbrot, B. Linguistique statistique macroscopique. In L. Apostel, B. Mandelbrot, and A. Morf. *Logique. langage et theorie de l'information*. Paris: Universitaries de France, 1957, 1-78.

Mandelbrot, B. A note on a class of skew distribution functions: analysis and critique of a paper by H.A. Simon. *Infor. and Control*, 1959, *2*, 90-99.

Mandelbrot, B. The Pareto-Levy law and the distribution of income. *Inter, Econ. Rev.*, 1960, *1*, 79-106.

Mandelbrot, B. Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon. *Infor. and Control*, 1961a, *4*, 198-216.

Mandelbrot, B. Post scriptum to "final note." *Infor. and Control*, 1961b, *4*, 300-304.

Mandelbrot, B. On the theory of word frequencies and on related Markovian models of discourse. In R. Jackson (Ed.) *Structure of language and its mathematical aspects*. The American Math. Society, 1961c, 190-219.

Mandelbrot, B. A survey of growth and diffusion models of the law of Pareto. *IBM Research Note NC-253*. 1963a, 1-39.

Mandelbrot, B. Word frequencies and the log-normal function. *IBM Research Note NC-265*, 1963b, 1-9.

Miller, G. A. Some psychological studies of grammar. *Amer. Psychol.*, 1962, *17*, 748-762.

Miller, G. A. and Chomsky, N. Finitary models of language users. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.) *Handbook of mathematical psychology.* Vol. 2. New York: Wiley, 1963, 419-491.

Miller, G. A. and Newman, E. B. Tests of a statistical explanation of the rank-frequency relation for words in written english. *Amer. J.* Psychol., 1958, LXXI, 209-218.

Miller, G. A., Newman, E. B., and Friedman, E. A. Lengthfrequency statistics for written english. *Infor. and Control*, 1958, *1*, 370-389.

Miller, G. A. and Selfridge, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, *63*, 176-185.

Miron, M. S and Wolfe, S. A cross-linguistic analysis of the response distributions of restricted word associations. *J. verbal learning and verbal behavior,* 1964, *3*, 376-384.

Myers, J. L. and Atkinson, R. C. Choice behavior and reward structure. *J. Math. Psychol.*, 1964, *1*, 170-203.

Osgood, C. E. On understanding and creating sentences. Urbana, Ill., 1964. Unpublished manuscript.

Osgood, C. E. and Miron, M. S. (Eds.) *Approaches to the study of aphasia*. Urbana, Ill.: Univer. of Ill. Press, 1963.

Pareto, V. *Cours d' economie politique*. Lausanne and Paris: Rouge, 1897.

Pretorius, S. J. Skew bivariate frequency surfaces, examined in the light of numerical illustrations. *Biometrika*, 1930, *22*, 109-223.

Rapoprt, A. Coment: the stochastic and the "teleological" rationales of certain distributions and the so-called principle of least effort. *Behav. Sci.*, 1957, *2*, 147-161.

Rapoport, A. Type-frequency distributions for speech of normal and aphasic speakers. Chapel Hill, N. C.: Psychometric Lab. Research Memo. no. 17, 1964.

Rapoport, A. and Horvath, W. J. A study of a large sociogram. Behav. Sci., 1961, 6, 279-291.

Rapoport, A. Comparison of four models for word-frequency distributions from normal and aphasic speakers. University of North Carolina, Doctoral Dissentation in Psychology, 1965.

Saporta, S. (Ed.) Psycholinguistics. New York: Holt, Rinehart and Winston, 1961.

Simon, H. A. On a class of skew distribution functions. Biometrika, 1955, 42, 425-440. (Also in Simon, H. A. Models of man. New York: Wiley, 1957, 145-164.)

Simon, H. A. Some further notes on a class of skew distribution functions. Infor. and Control, 1960, 3, 80-88.

Simon, H. A. Reply to "final note" by Benoit Mandelbrot. Infor. and Control, 1961a, 4, 217-223.

Simon, H. A. Reply to Dr. Mandelbrot's post scriptum. Infor. and Control, 1961b, 4, 305-308.

Simon, H. A. and Bonini, C. P. The size distribution of business firms. The Amer. Economic Rev., 1958, 48, 607-617.

Simon, H. A. and Van Wormer, T. A. Some Monte Carlo estimates of the Yule distribution. Behav. Sci., 1963, 8, 203-210.

Somers, H. H. Analyse mathematique du langage. Lois Generales et Mesures Statistiques. Vol. 1. Louvain, Belgium, 1959.

Stevens, S. S. Measurement, statistics, and the schemapiric view. Science, 161, 1968, 849-856.

Thompson, H. R. Truncated normal distributions. Biometrika, 1951, 38, 414-422.

Thorndike, E. L. On the number of words of any given frequency on use. Psychol. Rec., 1937, 1, 399-406.

Thorndike, E. L. and Lorge, I. A teacher's word book - 30,000 word list. New York: Teachers College, Columbia Univer., 1944.

Ullman, S. Semantics: an introduction to the science of meaning, New York: Barnes and Noble, 1962.

Wepman, J. M., Jones, L. V., Bock, R. D. and Van Pelt, D. Studies in aphasia: background and theoretical formulations. J. speech and hearing disorders, 1960, 25, 323-331.

Wepman, J. M. and Jones, L. V. The language modalities test for aphasia, Chicago: Univer. of Chicago, Ind. Rel. Ctr., 1961.

Williams, C. B. A note on the statistical analysis of sentence length as a criterion of literary style. Biometrika, 1940, 31, 356-361.

Willis, J.C. Age and area. Cambridge: The Univer. Press, 1922.

Wolberg, L.R. The technique of psychotherapy. New York: Grune and Stratton, 1954.

Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S.. Philos. Trans, of the Royal Society, Series B, 1924, 213, 21-87.

Yule, G. U. The statistical study of literary vocabulary. Cambridge: The Univer. Press, 1944.

Zipf, G. K. The psych-biology of language. Boston: Houghton Mifflin, 1935.

Zipf, G. K. Human behavior and the principle of least effort. Reading, Mass.: Addison-Wesley, 1949.

SECTION IV - BIBLIOGRAPHY

1. Aborn, Murray, and Rubenstein, Hubert Word class distribution in sentences of fixed length. Language, 1956, 32, 666-674.

> "Samples of printed English sentences of three lengths were drawn randomly from a representative selection of popular magazines. The words of each sentence were classified according to Fries's system and a count was made of the various word classes at each sentence position. The tabulations were plotted as frequency classes and treated statistically. Principally, the data obtained from all three sentence lengths indicate (1) that the greatest variations in word-class frequency tend to occur in sentence extremes and the immediately adjoining positions, and (2) that different word classes have characteristic patterns of variation."

2. Aborn, Murray, and Rubenstein, Hubert Perception of contextually dependent word-probabilities. American Journal of Psychology, 1958, 71, 420-422.

> "Eighty subjects were instructed to write down the first eight words coming to mind which could replace the missing word in a sentence, and then to rank these eight words in order of decreasing likelihood of occurrence in the sentence. The findings both for long and for short sentences may be summarized as follows: (1) words perceived as being more probable in a given context tended to be those actually occurring with greater probability in that context; and (2) greater agreements among subjects' responses were exhibited in the case of words perceived as more probable than in the case of words perceived as less probable. Together with the work of Zipf, these results suggest the following generalization: In contexts of low constraint, the number of different probabilities perceived is far less than the number of possible alternatives."

3. Allen, J. The Swahili and Arabic manuscripts on tapes. Leiden, The Netherlands: E. J. Brill, 1970.

> This is a compilation of unpublished literature in Swahili. It states the earliest found manuscript in Swahili is dated 1724. The scope is largely Swahili written in Arabic script. Part 1 is a serial list of holdings with descriptions; first those in Swahili and second those in Arabic. Part 2 is an Alphabetical list of Swahili manuscripts by titles and first lines. Tapes include verse and prose examples.

4. Allen, W. Living English structure (practice book for foreigners). London: Longmans, Green, and Company, Ltd., 1949.

   This book is an empirical approach using a series of 15 exercises which drill English structure into the student. The exercises are graded according to difficulty, as elementary, intermediate, and advanced.

5. American Mathematical Society Structure of language and its mathematical aspects. Proceedings of Symposia in Applied Mathematics, Providence, Rhode Island: American Mathematical Society, 1961 (2d Printing, 1964), 12.

   This is a compilation of 20 studies on the subject of linguistics, logic, and mathematics by well-known experts in these fields. Of particular interest is the chapter "On the Theory of Word Frequencies and on Related Markovian Models of Discourse" by Benoit Mandelbrot (pages 190-219). Mandelbrot's chapter (article) treats a variety of topics related to the models for the law of word frequencies by Estoup and Zipf. It discusses diachronic and synchronic aspects of the model. It also contains a criticism of certain attempts to apply lognormal probability distrubution to data on word frequencies. The final part is a discussion of linguistics and the role of statistical and other enumerational laws, such as the Willis Species-Genera relationships.

6. Ashen, R. Language--an enquiry into its meaning and function. New York: Hayes and Brothers Publishers, 1957.

   This is part of the Science and Culture Series. There are 19 chapters or articles by different authors, except that the first and 19th chapters are by the compiler himself. These two; "Language as Idea" and "Language as Communication", together with Chapter 9--Roman Jakobson on the Cardinal Dichotomy of Language--have relevance for statistical analyses or vocabulary.

7. Ashton, E. Swahili grammar (including intonation). London: Longman, Green and Company, Ltd., 1966 (13th Printing).

   This text is divided into two parts: Part 1, progressive lessons with exercises is concerned with everyday conversational phrases and literature and Part 2 which goes into more detail on each

7. (continued)

grammatical topic but has fewer exercises. There is a key to the exercises after the last chapter. The last chapter concisely summarizes the highlights of the book. There is a vocabulary of of nouns and verbs used in the lessons and exercises at the end of the book, followed by additional situational exercises; e.g., "at the office" and "at the hospital".

8. Bailey, D. Glossary of Japanese neologisms. Tucson: University of Arizona Press, 1962.

The purpose of this glossary was to collect in one place a list of useful new words and phrases not found in Japanese-English dictionaries, specifically Kenkyusha's New Japanese-English Dictionary of 1954. It includes proper nouns of considerable use, other useful words overlooked in the referenced dictionary, and a list of Japanese words in current use not found in Kenkyusha's new Japanese-English Dictionary. 12,000 candidate words were narrowed down to some 6,000. Sources of words are: "Basic Information on Current Words 1959-62", "Dictionary of Newspaper Terms 1960", "Handbook on Words in Current Use 1961", and "Dictionary of National Language 1960".

9. Bailey, Richard W., and Burton, Dolores M., S.N.D. English stylistics: a bibliography. Cambridge, Mass., and London: 1968.

A collection of over 2,000 items concerning general stylistics and style in English and American literature since 1500, the work is divided into three main sections: bibliographical sources, language and style before 1900 (including works on stylistics in antiquity), and English stylistics in the Twentieth Century.

10. Bakaya, R. M. An experiment in compiling a minimal vocabulary for reading scientific-technical literature in Russian. Babel, 13, 1967, 163-168.

The purpose is described as providing a minimal reading or receptive vocabulary for scientific-technical literature in Russian. The method was to compile the vocabulary from a comparative study of nine existing word lists, in general selecting those words which occur on at least three of the lists, resulting in a minimal vocabulary of some 3500 words which was then checked against three scientific texts with the result that the Bakaya list covered more than 95 percent of the texts directly or indirectly.

11. Baker, Sidney J. The pattern of language. *The Journal of General Psychology*, 1950, *42*, 25-66.

    After an extensive summary of previous investigations of lexical data both within and across linguistic boundaries, the author reports his calculations of word length (by letters) in several word lists. He concludes with a discussion of Zipf's law and polysemy.

12. Baker, Sidney J. Ontogenetic evidence of a corrleation between the form and frequency of use of words. *The Journal of General Psychology*, 1951, *44*, 235-251.

    Baker examined a 40,000 word collection of letters written by a paranoid schizophrenic and compared the rank-frequency distribution of words in the letters with similar lists published by Horn and Throndike.

13. Bar Hillel, Y. Logical syntax and semantics. *Language*, 1954, *30*. (Bobs-Merrill Reprint L-3.)

    A good part of this article is devoted to refuting Zellig Harris' (Methods of Structured Linguistics) contention that most considerations of meaning in linguistics can be satisfied by distributional procedures. Bar Hillel cites that most structural linguistics have recognized that not all aspects of linguistics can be handled by distributional analysis alone, in spite of Harris' thesis that he can explain synonymy and active-passive relationships. Bar Hillel attacks what he believes is Harris' basic assumption that "any two morphemes having different meaning will also differ somewhere in distribution." He says that by extension of this statement that many of the transformational aspects of language, if not all of them, could be reducible to the formational aspects. However, Bar Hillel says this is not true. He then proceeds to elaborate. If he dislikes Harris' thesis, he does agree with Rudolph Carnap who believes that logical analysis has an equal place with distributional analysis and that modern semantics *must* also be considered.

14. Barber, C.L. Some measurable characterisitcs of modern scientific prose. *Contributions to English syntax and philology*, ed. Frank Behre, Gothenburg: Adler, 1962, 21-43.

14. (continued)

The author is primarily concerned with identifying features of scientific language that will constitute particular difficulties for non-native speakers of English. He gives particular attention to clause and verb phrase structure and to the identification of words that appear frequently in scientific writing but not in the general vocabulary of English.

15. Barker, Muhammad Abd-al-Rahman, *An Urdu newspaper word count*. McGill University Institute of Islamic Studies, 1969.

"This volume is the last of four works dealing with the Urdu language prepared by the Institute of Islamic Studies, McGill University. The present volume, although not intended primarily as a dictionary, is suggested as a supplementary vocabulary source for further reading and research. The corpus upon which this work is based contains 136, 783 running words, collected from 15 Pakistani newspapers. The author's rules (which differ somewhat from those of Brill and Landau), as well as a discussion of word counts, the corpus of this work, word count methodology, Arabic orthography, and other pertinent information, are presented in the introductory section. Part One comprises the Urdu-English Alphabetical List, which gives the orthography, frequency, pronounciation, grammatical class membership, meaning, and usage of each lexeme. Part two, the Frequency List, relists all occurring words in descending order of frequency."

16. Barth, Gilbert *Rècherches sur la fréquence et la valeur des parties du discourse en français, en anglais, et en espagnol*. Paris: 1961.

A statistical study of the degree to which the three languages exploit the possible combinations of word classes.

17. Becker, Selwyn D., Bavelas, Alex, and Braden, Marcia An index to measure contingency of English sentences. *Language and Speech*, 1961, *4*, 139-145.

"Several indexes to measure contingency of sentences were constructed by considering nouns, repeated nouns, and total number of words. Contingency was operationally defined as reconstructibility in order to test the several indexes against a criterion. The best form of the index was then selected and retested. The contingency ranking, based on the index, of ten sections of text

17. (continued)

> correlated 0.84 with the reconstructibility ranking. It was concluded that the index is a valid initial approximation to a measure of contingency if contingency is defined as reconstructibility."

18. Becker, Selwyn D., and Carrol, Jean The effects of high and low sentence contingency on learning and attitudes. *Language and Speech*, 1963, *4*, 46-56.

> "By a logical analysis it was shown that the sentence contingency is roughly equivalent to Shannon's measure of redundancy. In two independent experiments it was demonstrated that a signficantly greater number of multiple choice questions are answered by those who study text characterized by higher sentence contingency, or redundancy. The findings were compared to those found in investigations of the effects of redundancy on words and syllables. Data from a third experiment provided support for the conclusion that preference for text material is also related to sentence contingency."

19. Beier, E. G., Starkweather, J. A., and Miller, D. E. Analysis of word frequencies in spoken language of children. *Language and Speech*, 1967, *10*, 217-227.

> The purpose of the study was to establish certain base rates in the language usage of children and to investigate some of the psychological significances of those base rates. The authors wanted to know whether their data would support Zipf's Laws, in particular whether in a given language sample, the number of different words would increase as the frequency of occurrence becomes smaller, and whether the magnitude of the words would tend to stand in inverse ratio to the number of occurrences of a given word. Additionally, the authors sought to determine which, if any, of a number of variables (such as the type/token ratio, word lists, magnitude of words, and the 10 most frequently used words) would differentiate the age groups. The experiment took place in Salt Lake City with grade school children. It is not clear what stimulus materials were used, but the boys were told not to use prepared speeches and taught how to handle tape recorders. Each boy recorded about 5000 words from which about 2700 were selected and compiled into two 40,000 word corpora for a grand total of 80,000 words for both groups. Five, one-

19. (continued)

minute samples of each boy's speech were recorded in order to obtain the rate of speaking in words-per-minute for each. The results were manipulated by an IBM 7094 computer and the two lists were compared by frequency between themselves and with the Eldredge Newspaper count of 1911. Each count had 42,000-43,000 words. The two children's counts each had about 3100 different words and the Eldredge count about twice that many (6000). The results tend to indicate that the printed language has a greater variety of expression than oral, which others have suspected. However, in this case, since printed adult news-writing was compared to two grades of elementary school oral output, some of the difference in variety is undoubtedly caused by the greater age and experience of the newswriters. In fact, differences between 6th and 10th graders tends to show somewhat greater variety among the 10th graders. The 1 percent level of significance , and the date indicate that older boys speak faster than younger ones, use significantly more positive and negative words, use slightly more singular self-reference, use slightly less plural self-reference, use more "other" references, and use slightly more "question" words. At equivalent intelligence levels, the two age groups showed insignificant differences in type/token ratios. Eight of the ten words in both boys lists were the same, although not in the same order. In the different words the 6th graders used "it" and "we", and the 10th graders used "not" and "do". In both groups, in general, shorter words tended to be used more often than longer ones. The authors hope to use their findings in developing a psycholinguistic profile of individuals for assessment of development, for developing reading materials better suited to age groups, for better understanding sequences in language development, and in inter-cultural comparison. A caveat is that with their small sampling (40,000 words) in one city (Salt Lake - 1965) far removed in time and space from the Eldredge sample (Buffalo - 1911) of only sex (male and of a different age group from the Eldredge count (grammer school vs adult) one must be careful about drawing sweeping conclusions.

20. Belevitch, Vitold On the statistical laws of linguistic distributions.

Annales de la sociétè scientifique de bruxeles, 1959, 73, 310-326.

"The rank-frequency diagrams of statistical linguistics are re-interpreted as distribution curves of the cumulative probability of types in the catalog versus the probability of tokens in the text. For such distributions, the closure condition $\Sigma p_i = 1$ (which does not hold in general statistics for the independent variable) imposes certain relations between the mean, the variance, the

20. (continued)

number of elements in the catalogue and the average information content (negative entropy). Sections 2 to 4 are devoted to the mathamatics of these relations, especially to their particular forms from truncated normal distributions. First and second order Taylor approximations to an arbitrary distribution law take the form of Zipf's and Mandelbrot's laws, respectively. Experimental data approximate a truncated normal distribution with $\sigma = 2.8$ bits as the general law for words. Data on letter and phoneme distributions seem to indicate that the standard deviation has a universal value of $\sigma = 1.4$ bits."

21. Belonogov, G. G. Raspredelinie castot pojavlenija flektivnyx klassov Rasskix slov. (Frequency distribution of the inflected word classes) (In Russian.) Problem Kibernetiki, 1964, 4, 189-198.

Statistical data concerning the distribution of inflected word classes in Russian were obtained by a computerized count from some half a million words of text.

22. Berckel, J. A., Van, Th. M., Corstius, H. Brandt, Mokken, R. J., and Wijngaarden, A. Van, Formal Properties of newspaper Dutch. Amsterdam: 1965.

Some 50,000 words were obtained from the issues of ten Dutch newspapers that appeared on June 19, 1959. In addition to examining the differences between the newspapers, the authors provide statistics for letter combinations, syllables, the rank-frequency relation of words, word length and type-token distribution of words.

23. Berger, K. W., An evaluation of the Thorndike and Lorge word count. Center States Speech Journal, 1971, 22 (1), 61-64.

"The publication by Thorndike and Lorge on the frequency of word appearance in English is often quoted as being representative of English speech. To examine possible differences in the word count by Thorndike and Lorge with contemporary printed materials a comparison was made between that work and a sample of 10,000 words taken from the November 20, 1969 issue of the "New York Times." The findings suggest a substantial but not dramatic difference between the two counts. Word comparisons from other contemporary printed sources would be useful, but researchers could concentrate their energy toward open word classes."

24. Berger, K. Conversational English of university students. Speech Monographs, 1967, 34(1), 65-73.

> "A study examining sentence length, phonetic content, word length, grammatical content, and word usage in student spontaneous speech. Sentences were collected and transcribed in informal settings. The average sentence was found to be 7.8 words with 23.5 phonemes. The most and least common phonemes are noted. The words "I" and "You" accounted for 7.2% of all words collected; 12 words comprised 25% of the words used; 50 words accounted for 46.5% of the conversations. Verbs appeared more frequently than any other part of speech followed by pronouns and nouns. Agreement in phonetic content and word frequency was found between these data and those of previous studies leading to the conclusion that these 2 parameters are reasonably stable in usage from late childhood through adulthood."

25. Berger, K., The most common words used in conversations, Journal of Communication Disorders. 1968, 1(3), 201-214.

> "Unguarded informal conversational vocabulary from a general adult population was sampled in the northeastern Ohio area. The sample produced 25,000 words of which there were 2,307 different words. A limited vocabulary usage and simple words as compared with more formal speech and with printed English. The words found in the present study are presented in an appendix. The appendix gives all of the words found, in alphabetical order, and includes variants of the base word where syllable length does not change. The usefulness and application of oral vocabulary as opposed to written vocabulary are briefly discussed. Further samplings of conversational speech, in spite of the difficulty as contrasted to printed materials, are recommended, particularly to determine consistency and variability based on geographical areas."

26. Berry, Jack Some statistical aspects of conversational speech Communication theory, ed. Willis Jackson, New York and London: 1953, 392-401.

> The article reports on an investigation of stress patterns in a 24, 781 word sample of conversational speech; the incidence of stress in high frequency function words is given partrticular attention.

27. Berry, Jack Oral data collecting and linguistics in Africa. Folklore Institute Journal, 6, 1969, 93-110.

27. (continued)

Discusses problems of selecting informants, eliciting and recording oral data. The article contains an appendix by Earl Stevick on the making and use of field tapes both for raw materials and as a basis for pedagogical treatment.

28. Black, John W., and Ausherman, Marian R. The vocabulary of college students in classroom speeches. Columbus, Ohio: The Ohio State University Bureau of Educational Research, 1955.

This study extends a prior study by Ausherman in 1950 entitled "Formal Spoken Vocabulary of College Students" and work done for the Office of Naval Research (ONR) by Kenyon College and Ohio University's Research Foundation. The informants were 274 male college students with samples taken from 607 classroom speeches. The objective was to obtain oral colloquial vocabulary in extemporaneous speech situations. The students were not typical college students, however. They were military enlisted personnel largely from the Midwest who were taking a background course in preparation for specializing in meteorology while in the service. They were highly intelligent, had high scholastic creditials, and high aptitude in mathematics. The samples were in general 3 1/2-4 minutes of speeches lasting five minutes on the average. The students used a microphone but thought it was a public address device since the recorder was in another room. Speeches had generally been outlined, but had not been written out or rehearsed. Recordings were transcribed for statistical analysis. Procedures for enumerating inflections followed Thorndike's procedures as far as possible. The corpus amounted to 288,152 running words including 6,826 different words. Frequencies ranged from 15,000 for "the" to nearly 2,000 words which occurred only once. Comparison of the statistics (oral-1955) with Thorndike's Teachers' Word Book of 20,000 Words (Printed) (either 1931 or 1944) was ambiguous. All of Thorndike's categories were represented in approximately the same relationship to each other as in his list, but were distributed differently in oral statistics. For example, Thorndike's first 1000 words accounted for only 14 percent of the first 1000 oral words (in order of frequency). In addition, 662 or nearly 10 percent of the oral vocabulary could not be found in Thorndike's 20,000 words. This is partly accounted for by the fact that many of the 662 words were neologisms, slang, occupational jargon, and colloquial compounds which employed non- and un- prefixes to form antonyms. Although groups of words in the speeches roughly corresponded to the same

28. (continued)

groups in written counts, there were many words in written lists that were not in the oral. Interestingly of Dewey's nine words making up 25 percent of English words used, the same words make up 22 percent of the oral list. Further, all the first 50 most common words in the oral list were found in Dewey's first 83 words and all but three (no, my, and me) of Dewey's first 50 words were found in the first 100 words. The authors note that two factors favorably affect the ability of a listener to understand oral language: familiarity (related to frequency) and number of syllables, with the former more important than the latter. They also note that oral vocabulary tends to be more restricted than written. The data from the study are presented in two lists: (1) A listing of words in descending order of frequency with breaks and summaries at selected frequency limits; e.g., 1000 and above, 100-999 and 50-99. (2) An alphabetical listing of words keyed to the related frequency groups in which the same words will be found in List 1.

29. Blankenship, Jane. A linguistic analysis of oral and written style. Quarterly Journal of Speech, 1962, 48, 419-422.

This study of four samples each of writing and formal speeches analyzed according to the method of C.C. Fries; the percentage of occurrence of each word class by position in the sentence and subcategories of the verb are studied. The author concludes that syntactic structure is more indicative of individual style than of the mode of discourse.

30. Bloch, B., and Jorden, E. Guides' manual for spoken Japanese, basic course, units 1-30. New York: Henry Holt and Company, 1946.

This book is almost entirely in Japanese. Section A includes basic sentences, pronunciation practice, practice in basic sentences, notes, exercises, check-up exams, and review of basics. Section B is the same as A for different basic sentences. Section C covers final check-up, listening in and free conversation. (Also published for the Armed Forces.)

31. Bongers, H. The history and principles of vocabulary control. Woerden, Holland: Wocopi, 1947.

The book was written in the context of teaching foreign languages in general and English in particular. While recognizing the problems of syntax or word usage for the person learning a language,

31. (continued)

this book concentrates only on vocabulary. The book consists of three parts: Part I is a general treatment of vocabulary including definitions, and history. It also includes Palmer's (Belgian) contributions, graded texts, quantitative statistics, classroom vocabularies, basic English (negative conclusion), world language and a comparison of several word lists: Thorndike--20,000, Faucett and Maki--6,000, Palmer--3,000, Palmer, Faucett and Wey--2,000, Palmer and Hornby--1,000, and Eaton--739. From a study of the above, the author derives a new 3,000 item word list. (The KLM List). Part 2 is a critical review of various word lists and includes thirteen appendices and a bibliography. Part 3 is a tabulation of the author's KLM List.

32. Booth, Andrew D. A 'law' of occurrences for words of low frequency. *Information and Control*, 1967, *10*, 386-393.

"The way in which the number of words occurring once, twice, three times, and so on in a text is related to the vocabulary of the author has been investigated. It is shown that a simple relationship holds under more general conditions than those implied by Zipf's law."

33. Borko, Harold (ed.) *Automated language processing: the state of the art*. New York: Wiley, 1967.

This is a collection of eleven original essays divided into three parts: "Language Data Processing," "Statistical Analysis," and "Syntactic Analysis."

34. Bourne, Charles P., and Ford, Donald F. A study of the statistics of letters in English words. *Information and Control*, 1961, *4*, 48-67.

"Data which had previously been published by several authors to describe the statistical characteristics of English words were examined to show the extent of their agreement. In addition, a detailed empirical study was made of two special types of English words: subject words and proper names. The statistical parameters which were measured and compared are: the distribution of letters, the distribution of terminal letters, the composite or total distribution of letters, the distribution of characters for each letter position, the distribution of characters for each letter position, the distribution of bigrams, and the distribution of word lengths."

35. Bowen, John H. Frequency stability of adjective trait names. *Psychological Reports*, 1972, *30*, 477-478.

"Using frequencies from the earlier Thorndike-Lorge and the later Kucera-Francis frequency counts, a lognormal distrbution model is applied to judge shifts in the frequencies of occurrence of trait adjectives from a likeableness scale. In the time between frequency counts, the frequencies of the adjectives shifted an average of approximately .68 words per million toward higher frequencies of occurrence. The amount of shift would probably not vitiate the generalizability of results based upon the Thorndike-Lorge count."

36. Brain, J.L. *Basic Structure of Swahili*. Syracuse, New York: Syracuse University Program of East African Studies, 1968.

This was an interim grammar of Swahili until a full reference grammar could be produced. It was written in East Africa as a teacher's guide and students' reference for an oral Swahili course. It is designed for the quicker coverage (two semesters) of the five semesters for the Foreign Service Institute Course. The lessons take up various aspects of basic grammar. There is a basic vocabulary and series of exercises with Swahili and English translations.

37. Brain, J.L. *A short dictionary of social sciences terms for Swahili speakers*. (Program for East African Studies, Occasional Paper #51) Syracuse, New York: Syracuse University, September, 1969.

The purpose of the dictionary is to provide Swahili speakers a vocabulary in the social sciences in the form of a dictionary. Terms were selected from UNESCO's "A Dictionary of the Social Sciences" by Gould and Kolb (ed.).

38. Brain, J.L. *Basic structure of Swahili, Part II (a background to the Swahili language and advanced exercises)*. (Syracuse University Program for East African Studies) Syracuse, New York: Syracuse University, August, 1969.

This booklet contains a brief background of Swahili in pages 1 to 19. The exercises (pages 21 to 34) provide practice in useful sentences and also provide the vocabulary to understand them.

39. Brain, J.L. *A social science vocabulary of Swahili*. (Program for East African Studies, Occasional Paper #33) Syracuse New York: Syracuse University, 1969.

The vocabulary is the beginning of the dictionary for personnel studying Swahili and Swahili areas. It is based on newspapers and political manifestos. It is arranged in Swahili-English ordering.

40. Buchanan, A., and MacPhee, E. *An annotated bibliography of modern language methodology*. (American and Canadian Committees on Modern Languages, Toronto, Canada: University of Toronto Press, 1928, 8.

This bibliography is arranged according to subject matter, such as: references, histories, aims and methods, learning processes, tests and examinations, texts used abroad, and miscellaneous. It is obviously dated.

41. Buchanan, M. *A graded Spanish word book*. (American and Canadian Committees on Modern Languages) Toronto, Canada: University of Toronto Press, 1927, 3.

In his introduction, Buchanan refers to earlier frequency studies in other languages: Kaeding in German, 1898, Thorndike in English, 1921, and Henmon in French, 1924. The purpose of preparing this frequency word list was to provide material for graded vocabulary tests, but it has become a standard, consistently used and referred to by later compilers of word lists in Spanish and other languages. The author took samples of 30,000 words each from 40 categories of printed material which were grouped under seven subject headings to obtain a total corpus of 1,200,000 running words. Subject headings included: plays, novels, verse, folklore, miscellaneous press, technical literature, and periodicals. Buchanan made the assumption (since recording devices were not well developed at the time), that an oral word count would not differ materially from his written one. Buchanan did recognize that what he developed was an "essential" word list which would have to be augmented in technical and specialized areas. To give weight to words which appear in many or most of the 40 categories, the number of categories was divided into the frequency and the quotient multiplied by 100 to give a credit number. The types were found to be 18,331 out of the 1,200,000 running words. 5,324 words had a frequency of 10 or more. Buchanan eliminated 189 words from his count as being too common: they do appear alphabetically in Part I of his

41. (contined)

list, however. Part 2 lists the total word count in order of frequency. These words appear 10 or more times (frequency of 10 or more), or they must appear in at least 5 of the 40 categories. Part 3, provides an alphabetical listing of the words giving their frequency, range, and merit number.

42. Buettner, C. Basic instruction in the Swahili language (self-instruction) Huelfsbuechlein fuer den ersten unterricht in der Suahili-sprache). Leipzig, C.D.: Weigel Nachfolger, 1891.

This book is in German. It is a booklet of grammar and exercises for the German speaker (reader) to use in learning basic Swahili. It updates some of Bishop Steer's work but it is obviously not current.

43. Bull, William E. Natural frequency and word counts. The Classical Journal, 1948, 44, 469-484.

The subtitle of this article truly represents its content: "The Fallacy of Frequencies". It is an extremely interesting study which helps explain some of the devices to which word-counters have had to resort in compiling their lists (e.g. Listing 50-150 most common word separately at the beginning of the count before assigning frequencies; addition of utility or available words (the concrete nouns) which carry situational meanings but, because situational or specific, have very limited frequencies in any limited sampling of printed, written or spoken language, and the problems of tapping the semantic or content-bearing words without which the lexical units convey no--or erroneous--communications. There appears to be an inverse relationship between natural frequency of parts of speech (i.e. the total number of individual words of a type such as noun or verb) at least in Indo-European languages, and the frequency with which these words are used, i.e. the greater the number of individual content-bearing words relating to specific items or situations; the less frequenctly they will each be used, whereas the lesser number of linguistically useful words, such as conjunctions, articles, prepositions, and relating verbs which tie the content-bearing words together are used over and over again regardless of the situation and thus generate a high statistical frequency out of proportion to their utility in learning a language. The author's summary is illuminating as to his points: (1) any word count is a statistically valid report only on what is included within it, (2) extremely high-frequency words are rarely the content-bearing elements of any communication, (3) range and frequency of words are determined by two sets of forces: linguistic and cultural, (4) it cannot be assumed that there is a correlation between frequency and utility, (5) word counts based on the hypothetical existence of the (any) language as a static entity cannot be considered a valid representation of a people's cultural and linguisitc activities and hence are of dubious value from a pedogogical point of view. The author's final indictment comes in his last

43. (continued)

paragraph: "From the foregoing evidence it would seem proper to draw the conclusion that there are so many factors and so many uncontrollable elements in life and language that no satisfactory results can be obtained by attempting to reduce natural heterogeneity to an artifical homogeneity by statistical methods. It may be concluded, although it is done so with considerable reluctance by the writer, that word counts cannot be considered a valid representation of a people's cultural and linguistic activities and that as a result their pedogogical usefulness is extremely dubious."

44. Bull, William E. Natural frequency and word counts. *Classical Journal*, 1949, *44*, 469-484.

"1. Any word count is a statistically valid report only on what is included in it. 2. Extremely high-frequency words are rarely the content-bearing elements of any communication. 3. Range and frequency of words are determined by two sets of forces: cultural and linguistic. 4. It cannot be assumed that there is a correlation between frequency and utility. 5. Word counts based on the hypothetical existence of the (any) language as a static entity cannot be considered a valid representation of a people's cultural and linguistic activities and hence are of dubious value from a pedagogical point of view."

45. Burton, N.G., and Licklider, J.C.R. Long-range constraints in the structure of printed English. *American Journal of Psychology*, 1955, *68*, 650-653.

"An experiment modeled after Shannon's was conducted to determine the extent to which estimates of the letter redundancy of English texts are dependent upon the number of preceding letters known to the subject. Data obtained indicate that, while the estimate of relative redundancy increases as knowledge of the foregoing text is extended from zero to approximately 32 letters, increasing the known number of letters beyond 32 does not result in any noticeable rise."

46. Bushnell, Paul F. *An analytical contrast of oral and written English*. New York: Teachers' College, 1930.

Various aspects of the language of student compositions are correlated with expert judgments of their merits; sentence- and word-level features are measured and "errors" of various kinds are tabulated.

47. Card, William, and McDavid, Virginia English words of very high frequency. *College English*, 1966, *27*, 596-604.

> The authors examine a variety of word frequency counts and discuss the biases inherent in them.

48. Carroll, John B. Diversity of vocabulary and the harmonic series law of word-frequency distribution. *The Psychological Record*, 1938, *2*, 379-386.

> The author is interested in a diversity equation whereby the relation of the number of different words in a vocabulary can be estimated despite variations in sample size. Illustrative material is provided from Santayana's, *The Last Puritan*, Hanley's word count of Joyce's, *Ulysses*, and the word-frequency lists compiled by Eldridge and Dewey.

49. Carroll, John B. How often a word? Review of John W. Black and Marian Auscherman's the vocabulary of college students in classroom speeches. *Contemporary Psychology*, Columbus, Ohio: Bureau of Educational Research, Ohio State University, 1946, 1, (7), 220.

> Carroll calls the Black and Auscherman count the most extensive oral one yet and welcomes it, since he believes the Thorndike word-count was not really representative because of its heavy emphasis on the Bible and older literary forms as opposed to contemporary sources. Carroll also believes that the new count will be helpful in controlling the word frequency factor in future experiments. In parting, Carroll questions rhetorically (and without answer) whether a spoken vocabulary is different from a written one.

50. Carroll, John B. The contributions of psychological theory and educational research to the teaching of foreign language. *Modern Language Journal*, 1965, *49*, 273-281.

> "This address, given at the international conference on modern foreign language teaching (Berlin, September 1964), presents a general discussion of the present scope, role, and potential use of research in foreign language teaching methodology, and maintains that the best research is that which is closely allied with

BEST COPY AVAILABLE

50. (continued)

> theory, and the hardest to apply is that which has direct bearing on classroom procedure. It points out the great scope for development in the theory of foreign language learning, citing favorably the work of mathematical learning theorists who have devised exact equations for the rate at which material is learned or forgotten. The need of forming an accurate theoretical comparison between the "audiolingual habit" and "cognitive code-learning" theories is discussed, such experiments being difficult to control since it is almost impossible to predict the exact techniques a student will employ and since the theoretical contrast has not been sufficiently well conceived. Neither method is based on modern theories of the psychology of language learning, and the discussion concludes with a critical comparison of the two, recommending a joining of audiolingual technique with some of the better elements of cognitive code-learning theory."

51. Carroll, John B. Review of G. Herdan's the advanced theory of language as choice and chance. American Scientist, New York: Springer-Verlag, 1966, 54. (4), 480A-481A.

> Carroll does not like this book. He says it is mainly a reprint, with some exceptions, of parts of Herdan's earlier works, including his earlier book with the simpler title of Language as a Choice and Chance. He does not see how the material can be called "advanced", that it is at best elementary and it in some cases indicates a retrogression from Herdan's earlier books. He concludes by saying that in spite of some provacative material, Herdan has revealed himself as behind the times in linguistics and cannot pass as a mathematician's linguist or a linguist's mathematician.

52. Carroll, John B. On sampling from a lognormal model of word-frequency distribution. Computational analysis of present-day American English, Providence, R.I.: 1967, 406-413.

> "In our investigations thus far we have not yet arrived at an efficient method for estimating the parameters of the theoretical population from the characteristics of a sample" (page 413). The attempt to determine such a model from the empirical data in the Brown Corpus is discussed.

53. Carroll, John B. An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for a standard frequency index (SFI). Computer Studies in the Humanities and Verbal Behavior, 1970, 3, 61-65.

> "A new word usage coefficient, U_m, is proposed. It is an adaptation of Juilland's U but in contrast to U it (1) can be computed from frequencies in unequally-sized categories, (2) uses a more appropriate measure of the dispersion of word probabilities over categories, (3) will not take the value zero when all occurrences are concentrated in a single category, and (4) is always scaled in terms of a corpus of a standard million tokens. Computations are given for illustrative data and discussed. For many purposes, however, a logarithmic frequency scale is more convenient and meaningful, and it is thus proposed that frequency data be scaled according to the formula SFI - 10 ($\log_{10}p$ + 10), where SFI is the Standard Frequency Index and p is the probability or proportional frequency of the item. An equivalent formula based on U_m is SFI = 10($\log_{10}U_m$ + 4). For most data from standard frequency counts, values of SFI will range from 35 to 90, each unit increment corresponding to an increase of about 25.9 percent in frequency."

54. Carroll, John B. Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 722-729.

> "Stevens' subjective magnitude estimation (SME) method was used in obtaining estimates of relative word frequency from two adult groups (15 lexicographers, 13 other adults) for 60 words ranging widely in objective frequency. Lexicographers rendered more reliable estimates, and their averaged data correlated more highly (.970) with objective log frequency than those of the second group (.923). The objective frequency of the first stimulus considered in the SME task is not related to an S's overall accuracy in predicting objective frequency, but accuracy is related to the S's tendency to perceive frequency ratios as relatively large. Subjective estimates measure available objective counts, and may be more valid measures of true word probability."

55. Carroll, John B. Current issues in psycholinguistics and second language teaching. Paper presented at the Fifth Annual TESOL Convention,

55. (continued)

New Orleans, La., March, 1971, Eric Accession No. ED-052-643.

> "Seemingly conflicting points of view concerning language instruction which are expressed in various teaching methodologies are reconciled in this paper. Key issues discussed include: (1) the nature of linguistic rules and their relation to the "habits" of language use, (2) the role of grammatical theory in language teaching, (3) the nature of language learning, (4) a balance between an audiolingual habit theory and a cognitive code theory, and (5) some of the critical variables in language pedogogy. The author illustrates why the field of language instruction has become characterized by pedagogical uncertainty and concludes that the teacher's ability to manage learning behavior remains one of the most unexplored, unstudied variables in educational research."

56. Carroll, John B. Behind the scenes in the making of a corpus-based dictionary and word frequency book. Paper presented at the meeting of the National Council of Teachers of English, Las Vegas, Nev., November, 1971, Eric Accession No. ED-056-842.

> The publication of the American Heritage Word Frequency Book and the American Heritage School Dictionary marked a new advance in the technology of dictionary and word-frequency book construction. The use of high speed computers enabled the compilers to analyze five million words from a body of materials frequently used in elementary and junior high schools. New mathematical techniques have improved the accuracy and scope of word frequency analysis. The word frequencies are listed by grades, thus enabling teachers and writers to get accurate information on the specific level they are interested in. References are included.

57. Carroll, John B., Davies, P., and Richman, B. Word frequency book. New York: Houghton-Mifflin Company and American Heritage Publishing Company, Inc., 1971.

> This is the most recent of vocabulary counts; and it is an excellent one, although it has limited application to adult spoken language since its samples were drawn from printed English to which children in grades 3 to 9 are exposed, with

57. (continued)

emphasis on grades 4 to 8. In addition, it concentrates on words rather than meanings so that the semantic part of word counts is not covered. The publication begins with a foreword and notes on the development of the corpus by Richman, notes on the Statistical Analysis by Carroll, and New Views on a Lexicon by Davies. The corpus was a computer-assembled selection of 5,088,721 words (tokens) drawn in 500 word samples from 1,045 published materials (texts and other student-used materials). It contains 86,741 different words (types). The materials from which the 500 word samples were drawn were textbooks, workbooks, student kits, novels. poetry, general non-fiction, encyclopedias, and magazines--as of November and December 1969. The samples reflect 22 subject areas, 17 of which were curriculum areas, three library categories, a magazine category, and a miscellaneous category which eventually turned out to be devoted principally to religion. The sampling of 1,045 texts was taken from 6,162 titles submitted in response to a national survey of U.S. schools, including public, Roman Catholic, and independent (private) schools, The 1,045 texts were in about 46 percent of the replies, although they constitute only about 16 percent of the 6,162 titles submitted. Machine processing of the data provided two types of output: citations--occurrences of types extracted in sufficient context to provide for the construction of definitions later forming the basis for the American Heritage School Dictionary--and descriptive statistics--frequency of occurrence and distribution. The statisitcal work is based on the lognormal model developed by Herdan. The results are displayed in alphabetical lists with frequencies indicated, classified by grade and by subjects. They are also displayed in frequency rank lists and frequency grouped distribution lists by total corpus, by grade, and by subject.

58. Carroll, John B., and Lamendella, John T. Subjective Estimates of Consonant Phoneme Frequencies. Educational Testing Service Research Bulletin RB-72-11, Princeton, New Jersey, 1972.

"Subjective magnitude estimates of the frequencies of 24 consonant phonemes were obtained from 65 university students, some with training in linguistics, by a method that had been used by Attneave (1953) for judgements of letter frequencies. Reliabilities of averaged judgements for comparably sized groups of 30 judges were estimated as in the neighborhood of .95. Averages of logarithmically transformed judgements were correlated with log frequencies from objective counts with coefficients in the range .736 to .876 (or .764 to .907 when corrected for attenuation). Despite the high reliabilities and predictive validities, there was evidence that the judgements were strongly influenced by experienced frequencies of letters of the alphabet."

59. Chaplin, H., Martin, S., and Nihonmatsu, R. Advanced Japanese Conversation. U.S. Department of H.E.W., Contract OE-3-14-005.

This book on advanced Japanese is designed as a follow-on to basic texts such as Jorden and Chaplin's Beginning Japanese which provides considerable facility in conversation. This book expands the conversational capability of the student by using three scenarios with a variety of realistic situations. Tape recordings were made by professional actors in Tokyo. Each scenario is backed up by vocabulary, notes and drill sentences. The sentences give practice in the grammatical points involved.

60. Chomsky, N. Logical syntax and semantics (their linguistic relevance). Language, January-March, 1955, 31, (1), (Part 1). (Bobs-Merrill Reprint L-3.)

In this paper, Chomsky comments on the Bar Hillel paper on the same subject. Basically, Chomsky takes issue with Bar Hillel's premises that logical syntax and semantics have disciplines or sub-disciplines which really furnish solutions to linguistic problems, especially those of transformation and semantics (as known at the time, i.e., 1955). Chomsky holds that they do not provide grounds for determining synonymy and consequent relations; they only point out that consequence is a relation between sentences, and synonymy a relation between words. Acknowledging that semantics is divided into a theory of reference and a theory of meaning, Chomsky states that Carnap's theory of meaning on which Bar Hillel bases his arguments is inadequate for linguistics. As for Bar Hillel's citation of M.V. Quine and Tarski in defense of "meaning", Chomsky says it is a mistake since their work was principally on the theory of reference which is of little use to linguists. Chomsky then takes issue with Carnap on the matter of models since Carnap believes artificial languages are necessary to the study of natural languages. Chomsky remains skeptical that a useful model can be constructed. Chomsky concludes by stating he believes Bar Hillel misunderstood Harris in his criticism of him in his article and then he objects to the thesis that incorporating logical syntax and semantics into linguistic theory will solve certain of its problems in that the theory of meaning in natural language is in any way clarified by constructing artificial languages in terms of rules which are called synonymous. Chomsky says we can solve the problems of synonymy and transformation in English in one of the following two ways, the latter being the better: by listing synonymous pairs under the heading "synonyms in grammar" and transformational pairs under the heading

60. (continued)

of "transformations" or by finding operational tests to determine their relationship and eliminate the need for arbitrary listings.

61. Chomsky, N. Review of 'Verbal Behavior' by B.F. Skinner. *Language,* January-March, 1959. (Bobs-Merril Reprint Series in the Social Sciences H-34.)

In his book, Dr. Skinner provides a functional analysis of "Verbal Behavior" in the context of his behaviorist psychology. In general, Chomsky disputes Skinner's claims, largely on the basis that Skinner's observations of the behavior of the lower animals cannot be applied in any really profound way to human behavior. Chomsky describes Skinner's concepts one by one and attempts to prove they do not describe verbal behavior if taken literally, or if taken metaphorically they do not add to current knowledge.

62. Chomsky, N. and Miller, G. Finitary models of language users. *Handbook of Mathematical Psychology (Chapter 13)*, New York: John Wiley and Sons, Inc., 1963, *2*, 419-491.

This chapter considers some of the models and measures that have been proposed to describe talkers and listeners, i.e., the users rather than the language itself. It is based on the fact that there is a distinction that a person's knowledge and his actual or potential behavior are not the same, so a formal characterization of a language is not at the same time a model of users. The authors state that in considering models for the actual performance of human talkers, an Important criterion of adequacy and validity is the extent to which the model's limitations correspond to actual human limitations. Two finite models are considered: the stochastic and the algebraic. The chapter concludes with a section on "Towards a Theory of Complicated Behavior". In constructing models, only the speaker-listener models were used instead of one for each. Stochastic theories of communications assume the array of message elements can be represented by a probability distribution and that communicative processes transform the probability distribution according to transitional probabilities. The section on Stochastic models contains a paragraph on word frequencies. With algebraic models, the purpose is to construct a model for the language user that incorporates a generative grammar as a fundamental component. This discussion

62. (continued)

concentrates on the listener and his faculties for perception, but only as a matter of convenience since the authors consider speaker-listener models as proper. **Preliminary evidence points** to the Chomsky idea of "kernel" or basic sentences that play a central role not only linguistically but psychologically as well, as the individual decides how to transform them into what he actually says (utterance) or understanding of what he has heard. In considering a theory of complicated behavior, the authors take into consideration among other things from linguistic theory: information and redundancy, degree of self-embedding, depth of postponment, structural complexity, and transformational complexity.

63. Chomsky, N. and Miller, G. Introduction to formal analysis of languages. Handbook of Mathematical Psychology (Chapter 11), Luce, R., Bush, R., and Galanter, E. eds. New York: John Wiley and Sons, Inc., 1963, 2.

In this study, Chomsky and Miller state that the fundamental fact that must be faced in any investigation of language and linguistic behavior is that a native speaker has the ability to comprehend an immense number of sentences he has never heard before, and to produce as the occasion requires, novel utterances that are understandable to other native speakers. In Chapter 11, they try to explain the following questions in elucidation of the statement above: What is the precise nature of the ability--the nature of language itself? How is the ability put to use, i.e., can we develop a model for users of a natural language? How is the ability developed in an individual? (Chomsky still rejects Skinner's characterization that language is a set of verbal responses.) Chomsky and Miller propose a theory of linguistic structure which must specify the class of possible sentences, the class of possible grammars, and the class of possible structural descriptions, and must provide a uniform and fixed method for assinging one or more structural descriptions to each sentence, generalized by an arbitrarily selected grammar of the specified form. The authors develop two conceptions of linguistic structure: a constituent structure grammar and the theory of transformational grammar. This book can **be seen as part of his continuing evolution of thought** on linguistics starting with his revolutionary "Syntactic Structure".

64. Chotlos, J. W. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 1944, *56*, 75-111.

"The present investigation is concerned with the relation of certain language variables to (1) the length of sample from which they are derived and (2) certain psychologically pertinent factors. In general, the language measures employed are based on a count of the number of different words (types) and the relationship of such measures to the total number of words, and to the factors of I.Q., chronological age, locality (city, town, rural), and sex..." A thousand samples of about 3,000 words each were collected from Iowa school-children over a five year period.

65. Chrètien, D. G. A new statistical approach to the study of language. *Romance Philology*, 1963, *16*, 290-301.

Review of Herdan's *Language as Choice and Chance*.

66. Cole, L. *The Teachers' Handbook of Technical Vocabulary*. Bloomington, Illinois: Public School Publishing Company, 1940.

This compilation draws on prior studies in various academic disciplines taught through high school level. The criteria used for vocabularies were frequency of occurrence, importance (according to experts), and social usefulness. The lists vary from 400 to 2000 in each of 13 subject areas, arranged in four groupings: mathematics (arithmetic, algebra, and geometry), language (English composition, American literature, and foreign language), social sciences (geography and history), and other sciences (hygiene, general science, chemistry, physics, and biology). The book is arranged with word lists broken down by grade level and includes a comparison with the Thorndike 20,000 word list. The author concludes that since no subject falls completely within the first 20,000 most common English words, some attention to vocabulary is required before any of the subjects can be taught effectively.

67. Condon, E. U. Statistics of vocabulary. *Science*, 1928, *67*, no. 1733, 300.

A discussion of the rank-frequency distribution of words in a text and proposed means for determining the mathematical law underlying the distribution. Carroll objects to this proposal since it makes diversity a function of sample size.

68. Daiji, S. *Japanese idioms (Nipongo no idiom)*. Tokyo: Sanseido, 1950.

This book distinguishes between idioms and free word combinations, and between the meanings of the idioms and those of the words which comprise them.

69. Dale, E., and Razik, T. *Bibliography of Vocabulary Studies*. Columbus: Ohio State University Bureau of Educational Research, 1963.

Contains 3,125 titles adding 542 new titles to the first edition published in 1957. References are arranged under 26 categories without annotation. It contains an author index.

70. Dale., E., and Reichert, D. *Bibliography of Vocabulary Studies*. Columbus: Ohio State University Bureau of Educational Research, 1957.

The First Edition of the Ohio State Bibliographic Project superceded by the 1963 revision prepared by Dale and Razik.

71. Davies, A. (ed.) *Language Testing Symposium - A Psycholinguistic Approach*. London: Oxford University Press, 1968.

This is a compilation of articles and studies on the language testing, including an introduction and one chapter authored by the compiler, and an appendix on item analysis. There are 11 articles in all, including the introduction, but excluding the appendix. The introduction covers language learning views and their influence on language testing, the uses of language tests, evaluation in language testing, language test analysis and Dr. Lado's approach to language testing. There are four main sections or groupings to the book after the introduction, although they are not listed as such: first section--evaluation, linguistics, and psychology (the basic disciplines and their relevance to language testing, chapters 2 to 4), second section--users and types of tests, chapters 5 to 8, third section--the influence of tests on education, chapters 9 to 11, and fourth section item analysis (the appendix). Two chapters are particularly relevant to spoken language teaching: chapter 7 on testing spoken language; some unsolved problems (G. E. Perren) and chapter 8 on the testing of oracy (skill in spoken language) (A. Wilkinson).

72 Denes, P.B. On the statistics of spoken English. *Journal of the Acoustical Society*, 1963, *35*, 892-904.

> "A variety of statistical information about spoken English was obtained. The data are the results of analyzing a considerable body of conversational material and narrative taken from 'Phonetic Readers'; the analyses were carried out by using a digital computer. The principles for selecting the speech material are discussed. Counts were obtained for the frequency of occurrence of phonemes, for the diagram frequencies of phonemes, for word length, etc. Stress was taken into consideration, and many of the statistics were obtained separately for stressed and unstressed syllables. In addition, the frequency distribution of minimal pairs was obtained. Minimal pairs are the phoneme pairs that minimally distinguish one word from another. All results were evaluated from the articulatory point of view. It was found that, in spoken English, dental and alveolar articulations predominate and that manner rather than place of articulation is the dimension that carries by far the greatest functional load."

73. DeVito, Joseph A. Comprehension factors in oral and written discourse of skilled communicators. *Speech Monographs*, 1965, *33*, 124-128.

> DeVito describes his work as "an attempt to compare written and oral samples of the work of skilled communicators for (1) overall comprehensibility as measured by close procedure and (2) significant differences in selected elements supposedly related to ease of comprehension." Item two includes vocabulary measures of difficulty and diversity, sentence-level measures, and an examination of "density of ideas."

74. DeVito, Joseph A. Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs*, 1966, *33*, 73-76.

> "The concern of the present study, based on 18,000 words of oral and written discourse by skilled communicators, was with six psychogrammatical factors. Oral language was found to contain significantly more self-reference terms, pseudo-qualify terms, allness terms, qualification terms, and terms indicative of consciousness of projection than written language."

75. DeVito, Joseph A. Levels of abstraction in spoken and written language. The Journal of Communication, 1967, 17, 354-361.

> "Samples of 8,000 words of oral and 8,000 words of written discourse, obtained from speech professors who had written extensively, were analyzed for the relative levels of abstraction. Oral language was found to be significantly less abstract and contained more finite verbs and fewer nouns of abstraction than written language."

76. DeVito, Joseph A. A linguistic analysis of spoken and written language. Central States Speech Journal, 1967, 81-85.

> "Samples of spoken and written language obtained from professors of speech who had written extensively were analyzed for the frequency of the four major parts of speech and for two grammatical ratios which measure degree of qualification. Five of the six measures employed differentiated the two forms of discourse at statistically significant levels." The measure that failed to discriminate the two modes was the noun-verb to adjective-adverb ratio.

77. Dewey, G. Relative Frequency of English Speech Sounds. Cambridge, Mass.: Harvard University Press, 1923 (1950 revision).

> Dewey's analysis of the relative frequency of English speech sounds was intended for application to shorthand, acoustic devices such as the telephone, and phonographs, and to the study of language change, history, and trends. This book investigates not only sounds (syllables) but also combinations of sounds (words). It contains a discussion of previous works in the field of quantitative analysis. The data base consisted of samples of written text drawn from newspapers, correspondence, novels, and other prose sources. Analysis of results revealed that: 9 words constitute 25 percent of the 100,000 running words (Corpus), 69 words constitute 50 percent of the 100,000 732 words constitute 75 percent of the 100,000, and 1027 words recurred more than 10 times in the 100,000.

78. Dingwell, W. Transformational and Generative Grammar - A Bibliography. Washington, D.C.: Center for Applied Linguistics, 1965.

> The purpose was to compile as complete a bibliography as possible of linguistic rules that relate to sentences. There are two

78. (continued)

principal parts to this bibliography: published works: books and articles, and unpublished works: conference papers. The lists are principally the works of the following schools of transformational grammar: Z. S. Harris (University of Pennsylvania), N. A. Chomsky (MIT), R. E. Longacre (Summer Institute of Linguistics), and S. K. Shaumyan (USSR).

79. Dixson, R. (ed.) *Las 2000 palabras mas usadas con mas frecuencias en Ingles (The 2000 most used words with the greatest frequency in English)*. New York: Latin American Institute Press, Inc., 1956.

The first 1,000 words follow the Thorndike-Lorge list. The second 1,000 words follow the Interim Report on Vocabulary Selection for teaching English as a foreign language (Palmer, Thorndike, West, Sapir, *et al.*) modified by current American-English usage. The words list is arranged alphabetically within groupings of 1 to 500, 501 to 1,000, and 1,001 to 2,000.

80. Dolby, J. L., Resnikoff, H. L., and MacMurray, E. A tape dictionary for linguistic experiments. *Proceedings of the Fall Joint Computer Conference 1963*, Baltimore and London: 1963, 419-423.

"A tape dictionary of some 75,000 entries has been prepared with part-of-speech, status, usage, graphemic syllabification and stress information. The entries have been sorted alphabetically forward and backward as well as by syllable and by part of speech. Comparisons are being drawn between various measures of usage as well as between to measures of the number of syllables in the written form. Considerable care has been taken to minimize the number of errors in the list and to insure a high degree of consistency in the coding. The authors believe that the resulting listing will be of great utility in basic studies of the nature of linguistic data handling." The project resulted in the production of: *The English Word Speculum*. 5 vols. Sunnyvale, Calif.: Lockheed Missiles and Space Company, 1964.

81. Driemann, G. H. J. Differences between written and spoken language: an exploration study. *Acta Psychologica*, 1963, *20*, 36-57 and 78-100.

The quantitative measures employed in this study include the total number of words in each sample, a classification of words

81. (continued)

by number of syllables, the verb-adjective ratio, and the type-token ratio. Texts from the writing and speech of eight psychology students were studied.

82. Durr, William K. A computer study of high-frequency words in popular trade juveniles. A paper presented to The International Reading Association, Anaheim, Cal., May 6-9, 1970.

"Word frequency was determined for library books that primary-grade children selected for free reading. A survey of librarains determined which books these children selected. This list was reduced to 80 books through evaluations by elementary school teachers. A computer analysis of each word in these books revealed 105,280 running words. When proper names, onomatopoeic words, and easily recognizable inflected forms and compounds were omitted, there were only 3,220 different words in all of these books. A frequency count of these different words revealed that just 10 words account for almost one-fourth of all running words, 25 words account for over one-third of all running words, and 188 words account for almost seven out of 10 of all running words. It was suggested that systematic teaching of these high-frequency words help insure that children have the background needed to read library materials of their own choosing at an early age. References and tables are given."

83. Eastman, Carol M. The status of the reversive extension in modern Kenya coastal Swahili. Journal of African Language, 1969, 8, (Part 1), 29-39.

This study is a fallout of a study on the Vumba, Amu, Bajuni and Jomvu dialects of Swahili conducted by the author in 1965-66. The procedure was to gather data in two hour sessions. At least two informants were questioned individually for each dialect with respect to 51 verbs. For each one they were asked to supply sentences exemplifying its use. The 51 verbs were common ones having the largest number of extended forms (The Standard Swahili-English Dictionary, Oxford, 1939 was used). In a later phase of the study, verbs with commonly occurring radical final elements were extracted from the dictionary. Such radicals fell into the category of compound radicals.

84. Eastman, Carol M. Markers in English-influenced by Swahili Conversation. (Papers in International Studies: African Series No. 8) Athens, Ohio: Ohio University Center for International Studies: African Program, 1970.

> This study examines one facet of changes in Swahili, the national language of Kenya and Tanzania (Tanganyika). It examines the use of markers as features of interference in Swahili-English bilingual conversation. These features involve the adoption of syntactic and semantic deviations in one language which can be attributed to the other. This paper demonstrates clearly how foreign words can be integrated into a language (Swahili) in an area where many people are bilingual (English-Swahili). Interference consists of simplification, lexical insertion of English words, English syntax incorporation, correction code switching (saying **the same thing in both languages), and improper** marker usage, i.e., as **transitional utterances and oral pauses.** Informants were two Tanzanians studying at a US university who were asked to talk about a variety of subjects as if they were in their own country. They used Swahili basically, but with considerable English intermixed. They actually talked on eight basic subjects. There were 503 utterances of varying length. Data were manipulated using a Burroughs 5500 Computer. The markers used were "nde", "nanhii", "kuma", and "tusema". (Markers are meaningless, so-called verbal pauses, hesitation words, such as "you know".) The study has an appendix containing English words used in the conversations.

85. Eaton, Helen S. Semantic frequency list for English, French, German, and Spanish: a correlation of the first six thousand words in four single-language frequency lists. Chicago: 1949, and New York: Dover Publications, Inc., 1961.

> This book is an extension of earlier single language word counts across language divisions in an attempt to correlate the word frequencies of a group of languages in order to show an interlingual relationship among the concepts measurable by a scale of frequency of use. In a more specific sense, it is an attempt to establish through relative frequency of use the common conceptions of mankind as they find expression in its various languages; in this case, four West and Central European languages, many of which have spread worldwide. Words have both form and meaning with

85. (continued)

meaning having the greater variation within any given language. Word counts, especially earlier ones, have tended to omit meaning (semantic value) and to concentrate on isolating the word forms most frequently used. Since a vocabulary of 500 words may represent a true vocabulary of 1500 or 2000 word meanings, knowing the form and a single meaning may not teach more than a small part of the living language. Dr. Eaton has tried to solve part of this problem in four languages with her comparative count which includes a semantic, as well as form count of some 6,000 basic concepts. The book is divided into an introduction, notes for the reader, and seven parts, one for each 1000 of the first 6000 words, and one for the first part of the seven **thousand**. These are followed by indexes to each of the word lists, an index of words deleted from prior English and German lists, and those moved from one group of frequencies to another (Appendix I), and a conceptual analysis of substantives, verbs, and adjectives in the lists (Appendix II). Sources of the word lists were Thorndike's "Teachers' Word Book of 20,000 Words," Vander Beke's "French Word Book.' (6,000 words), Kaeding's "Frequency Dictionary of the German Language" (80,000 words) and Buchanan's "Graded Spanish Word Book.' A great advantage of this book is the careful recording of procedures and sources used.

86. Edmundson, H.P. A statistician's view of linguistic models and language data processing. *Natural Language and the Computer*, ed. Paul L. Garvin, New York: McGraw-Hill, 1963, 151-179.

A survey of mathematical models of linguistic features.

87. Elderton, W.P. A few statistics on the length of English words. *Journal of the Royal Statistical Society*, 1949, *62*, 436-445.

This study examines a wide range of data in an attempt to determine the underlying laws of word-length distribution; Carlyle, Macaulay, Bacon, Scott, Swinburne, Johnson, Gibbon, Shakespeare, and The Bible provide the data. Further information is provided on the internal make-up of English words, including vowel and consonant distribution.

88. Eldridge, R. C. *Six thousand common English words, their comparative frequency, and what can be done with them*. Niagara Falls, N.Y.: 1911.

88. (continued)

A word count from newspaper prose; a sample of 43,989 words yielded 6,002 types.

89. Ellegård, Alvar Statistical measurement of linguistic relationship. *Language*, 1955, *35*, 151-156.

"Linguistic relationship has been measured statistically by means of the product-moment correlation coefficient, *r*. The linguistic meaning of various forms of this coefficient is discussed on the basis of a simplified model. It is maintained that the most satisfactory statistic measures degree of correspondence or similarity rather than relationship in the genetic sense. When applied to Indo-European data, the statistic results in good agreement with common philological judgement. Problems of significance are discussed. Finally it is concluded that the statistical technique will both require and help to establish a taxonomy of languages."

90. Ellegård, Alvar Notes on the use of statistical methods in the studies of name vocabularies. *Studia Neophilogica*, 1958, *30*, 214-231.

This article discusses various statistical methods for describing the distribution of personal names in a given area and concludes that some common techniques cannot be employed with curtailed samples or used for comparing name populations of different sizes. He suggests that his remarks on name vocabularies apply to vocabulary studies in general.

91. Ellegård, Alvar Estimating vocabulary size. *Word*, 1960, *16*, 219-244.

A discussion of the problems of determining vocabulary size from text samples.

92. Ellegård, Alvar *English, Latin, and morphemic analysis*. Gotenborg, Sweden: Elanders Boktryckeri Aktiebolog, 1967.

This is a short discussion and analysis of Latin root words, inflections, prefixes, and suffixes in English, and the derivation of seven rules for recognition of morphemic elements in English, either in words or separately.

93. Estoup, J. B. *Gammes stenographiques*. Paris: 1916 (4th edition).

An early attempt to **specify** the rank-frequency relation of words in a text.

94. Eyestone, Maynard M. Subordinate clauses in spoken and written American English. *Dissertation Abstracts*, 1967, *27*, 3857A.

The author analyses clause types and discusses them in a study of "...**50,000 words of unrehearsed comments by American journalists** and a like corpus taken at random from the published works of the same people."

95. Fairbanks, Helen The quantitative differentiation of samples of spoken English. *Psychological Monographs*, 1944, *56*, 19-38.

Three-thousand word samples were taken from ten "superior" college freshmen and ten schizophrenics (whose case histories are described). The speech of each subject was recorded and transcribed by the author. Comparative data is provided with particular attention to type-token ratios, grammatical structures, and word frequencies.

96. Flood, W., and West, **M.** *Dictionary of Scientific and Technical Terms*. London: Longmans, Green and Company, Ltd., 1960. (2d edition).

This dictionary contains 10,000 scientific and technical words for the layman on 50 subjects. It explains the words with a vocabulary of 2,000 words; 56 of which are technical and 120 more which may be difficult for children or individuals who are not native English speakers. However, most of the 120 words are explained in the dictionary itself.

97. Fowler, Marray Herdan's statistical parameter and the frequency of English phonemes. *Studies presented to Joshua Whatmough*, ed. Ernest Pulgram, 's-Gravenhage: 1957, 47-52.

This article examines the usefulness of Herdan's "coefficient of variation for the sampling distribution of means" in an examination of phoneme distribution in works by Graham Greene, Carl P. Boyer (a calculus textbook), and Beatrix Potter.

98. Franklin, H., Meikle, H., and Strain, J. Vocabulary in Context. English Language Institute, University of Michigan Press, 1964.

> This book presents vocabulary in context in the order of attention pointers (lexical area), presentation (conversations in context), generalization (explanation or notes on the conversations), and practice (drill exercises in situational contexts).

99. French Ministry of National Education Fundamental French/(first level) (Le Français fondamental (1er degré). Paris, France: National Pedagogical Institute, 1959 (2d edition).

> Fundamental French (1st Level) replaces Elementary French (1954) which was created in response to a request by UNESCO in 1947 for a daily spoken language to enlarge the worldwide education base, and in response to a need felt by the French to compete with Basic English, while not imposing restrictions on growth which are inherent in Basic English. Fundamental French (1st Level) is meant to be the basis for textbooks on French vocabulary and grammar to be taught to foreigners as their first real introduction to the French language. It is based on French spoken in as natural a situation as possible. 163 conversations were recorded in the Paris area from a wide range of persons, for a total of 312,135 running words (tokens) of which 7,995 were different (types). From this base, a frequency list was prepared. It was found that some very useful words were used relatively infrequently in both spoken and written French. They were usually concrete words such as bus, stamps, and grocer. To avoid losing them, it was decided to classify by the term "Availability", as well as "frequency". Words were listed by alphaabetical order and then grouped by meaning. The lists were then cut by 100 to a figure of 800 words indicated by frequency as valuable because they were close synonyms of others on the list, vulgar words, or presented some difficulty of use or learning. Some words were added to ensure all essential educative concepts would have a means of expression. The list has 1445 items of which 1176 are lexical and 269 are grammatical. For the grammar, constructions rare in spoken French were eliminated, such as some verb forms, interrogative expressions, and little used grammatical words. Both vocabulary and grammar are arranged by essential words and those of secondary priority. The vocabulary list is arranged

99. (continued)

In alphabetical order. Additionally, grammatical words, numbers, days of the week, months, seasons, and terms denoting human relationships are grouped separately. Frequency does not appear in the list in order not to influence teachers unduly. For certain classes such as vegetables, fruits, domesticated animals, metals, tools, and construction materials, only a minimum list was provided to allow for regional additions as required.

100. French Ministry of National Education *Le Francais fondamental (deux degre).(Fundamental French, 2d level (stage))*. (Brochure 707, E./SR) Paris: National Pedagogical Institute, (1963).

This book extends the 1st stage of Fundamental French for those who desire to acquire a more complete knowledge of the French language and culture. It is based essentially on the written language enriched by more precise grammatical words and is able to express thoughts with greater consideration for the affective and cognitive nuances. It corresponds to the essential needs of the real world. This second stage is designed to assist the learner to read books, newspapers, and periodicals. The vocabulary includes words from the word list produced for the 1st level with frequencies equal to or greater than 20 (the 1st level included words only down the frequency scale as far as 29.) It also includes words eliminated in the 1st level. It includes the remainder of the "available" words not included in the 1st stage. It includes new research--new study of varied types of printed texts to update the Vander Beke Dictionary, (Vander Beke's French Frequency Dictionary words with frequencies of 60 or above of 1,147,748 running words, but based on written text...and old--about 1900), resulting in 425 units of 500 words each not in the 1st stage, retaining those with a frequency of 13 or greater; study of a terminal textbook on education in civics for the last part of primary education, retaining words with a frequency equal to or greater than 7; study of psychological vocabulary based on studies of 160 students at eight normal schools, retaining words used by at least 15 of the 160; additional words not indicated by frequency but judged by experts to be required. It has an alphabetically listed vocabulary. The grammar is extended beyond that of the 1st stage to include constructions required to read written material, but still not a complete French grammar. It distinguishes (in the vocabulary) between words required for active use and those required only for understanding words when read or heard.

101. French, Norman R., Carter, Jr., Charles W., and Koenig, Jr., Walter The words and sounds of telephone conversations. *The Bell System Technical Journal*, 1930, *11*, 290-324.

"This paper presents data concerning the vocabulary and the relative frequency of occurrence of the speech sounds of telephone conversation. Tables are given showing the most frequently used words, the syllabic structure of the words, the relative occurrences of the sounds, and, for each vowel, the percentage distribution of the consonants which precede and follow it. Comparisons are made with the vocabulary and relative occurrence of speech sounds in written English."

102. Fries, C. *The Structure of English*. New York: Harcourt, Brace, and Company, Inc., 1952.

This book is a continuation, extension, and expansion of Fries' "American English Grammar" (English Monograph No. 10 of the National Council of Teachers of English, 1940), with respect to the sentence. The materials for analysis were more than 250,000 running words of standard English conversation recorded mechanically in Ann Arbor, Michigan. There were some 300 informants who provided about 50 hours of diverse conversation. Emphasis is on the grammar of structure of oral English, as opposed to the grammar of usage based on differences of writing of socio-economic classes. This has led to the identification of patterns of oral English. Unfortunately, Dr. Fries has dwelt more on the analysis of his findings than on his procedures in arriving at them. The last chapter, 13, on practical applications has much that is of use in the teaching of English to those for whom it is not the native tongue.

103. Fries, C. S., and Fries, A. C. *Foundations of English Teaching.* (The English Language Exploratory Committee) Tokyo, Japan: Kenkyusha, Ltd., 1961.

This book is one which provides a basis for building textbooks and teacher's guides for teaching English in Japan, especially to the first three grades of the lower secondary schools in Japan. It contains structures (patterns) and vocabulary. It emphasizes dialogue as the form of teaching; i.e., of the structure and pattern of word usage in English. The essence of the procedure for vocabulary selection is in chapter 1. (The Nature and Function of a Corpus; with corpus being defined as vocabulary and structure of social situational frames.) It does not give frequency counts, but supplies basic vocabulary and situational context for the use of words of the vocabulary.

104. Fries, C., and Traver, A. *English Word Lists - A Study of Their Adaptability for Instruction*. Ann Arbor, Michigan: the George Woke Publishing Company, 1950 and 1965.

> This work discusses English word lists, vocabularies, and the procedures for selection of vocabulary (frequency count, minimum basic lists, or psychological criteria). It is a critical inquiry into the character of the lists and their applicability to teaching English to non-English speaking learners. Specific discussions are included on the following: Ogden--basic English, West--definition vocabulary, Palmer and Hornby (IRET)--standard English vocabulary, Thorndike--teachers' word books, Faucett, Palmer, West, and Thorndike--interim report of vocabulary selection, Faucett and Maki--1534 words and values of 1 to 34, and Aiden--little English.

105. Frumkina, F. M. *Statisticeskie metody izucenija leksiki (Statistical methods of vocabulary study)*. Moscow: 1964.

> The author discusses both general problems and proposed models (e.g., Zipf's "law" of the statistical properties of the lexical structure of texts). Procedures for compiling a frequency dictionary are described. The text includes an appendix listing the most frequently used words in Puskin's lexicon and the statistical properties of Puskin's texts are given with particular attention devoted to the type-token relationship.

106. Frumkina, F. M. Allegemine probleme der haeufigkeitswoerterbuecher. *IRAL*, 1964, *2*, 236-247.

> The author reviews the (then) recent word counts of Garcia-Hoz and of Josselson, and proposes a method based on the Zipf function which will make it possible to compile a list with precision about a given percentage of the words in a text. First, a numerical estimate is made of the lowest frequency that in any particular list can be reached within a predetermined margin of error, then the size of the corpus is calculated which is necessary to determine the given frequency within the stated margin of error. Ms. Frumkin concludes with a list compiled according to her method which demonstrates its usefulness.

107. Frumkina, M., A.P. Vasilevich and Y.N. Gerganov, Sub'ektivnye otsenki chastot elementov teksta i zritelnoe vospriyantie rechevoi informatsii (Subjective estimates of the frequencies of textual elements and the visual reception of spoken information). *Nauchno-Tekhn. Infor. Prots. Sist.*, 1970, 9, 20-24. (In Russian)

"A discussion of the results of an experimental testing of the following hypothese: (1) the occurrence probability of meaningless letter combinations in speech predicts the threeshold of their visual recognition; and (2) subjective estimates of probabilities of letter combinations as obtained by psychometric techniques are a stronger predictive factor than the estimates of the same probabilites obtained by text counts. Tests were made using Russian trigrams presented tachistoscopically. The results justify the assumption that prediction of an individual's behavior in a new situation is based on subjective estimates of probabilities of the situation structures."

108. Fry, Dennis The Frequency of Occurrence of Speech Sounds in Southern English. *Archives neerlandaises de phonetique experimentale*, 1947, 20, 103-106.

Fry examined a corpus of 17,000 sounds of southern American English and provided frequency data--based on the transcription system formulated by Daniel Jones.

109. Fucks, Wilhelm On the mathematical analysis of style. *Biometrika*, 1952, 39, 122-129.

"Every significant text of a grammatical exposition consists of a certain material, the vocabulary, and some structural properties, the style, of its author. The passive vocabulary is formed by the totality of all words of that language, s, the author writes in, the active vocabulary is formed by a certain set, s', of that totality, the selection of which is determined essentially by the sort of literature the text belongs to and depends only in a lower degree on the peculiarity of the author. Style, however, is characteristic of the author at a certain period of his personal development. The aim of the following investigation is to formulate mathematically some of the peoperties of structure constituting style, so that for a given text the application of a simple mathematical criterion allows its attribution to a particular author at a certain period of his mental development."

110. Fucks, Wilhelm Mathematical theory of word formation. Information Theory, ed. Colin Cherry, London: 1956, 154-170.

The author hopes to discover "whether the process of word formation out of syllables in literary texts obeys a law which can be given mathematically." Word-length data from texts by Shakespeare, Aldous Huxley, Sallust, and Caesar is considered.

111. Gammon, Edward R. A statistical study of English syntax. Proceedings of the Ninth International Congress of Linguists, ed. Horace G. Lunt, The Hague: 1964, 37-43.

"This paper summarizes a statistical approach to English syntax. We show a segmentation of utterances based on the estimated sequence of forms of an utterance. We require that segment boundaries occur at positions in the sequence where the uncertainty in predicting possible future forms, given one or more immediate forms, is high. By 'high' we mean either in a relative sense, or larger than some prespecified value. The segments obtained from sequences of distribution classes coincide with recognizable phrases. Using various systems of phrases labeling, predictability of phrase types yields recognizable clauses and sentences; although these do not necessarily coincide with intonation patterns indicated by punctuation."

112. Garcia Hoz, V. Vocabulario usual, vocabulario commun, y vocabulario fundamental (Usual vocabulary, common vocabulary, and fundamental basic vocabulary). Madrid: Consego Superior de Investigaciones Cientifices (Supreme Council for Scientific Investigations, Institute San Jose de Calasaz), 1953.

An interesting distinction is made between active vocabulary used for speaking or writing and a passive or latent vocabulary used for word recognition as in reading or listening. The sources of the common vocabulary were private letters, periodicals, official, religious, and trade union documents, and books. Note that they are all printed or written sources. The book goes into considerable philosophical discussion on the number of words in the corpus to get a fair sample of different words for each type of source of words and in the determination of what constitutes a "word". As finally decided on, the usual vocabulary includes 12,428 words,

112. (continued)

arranged in alphabetical order with frequencies listed for each of the four sources as well as a total frequency. The usual vocabulary is supplemented by 369 words of restricted usage--mainly technical and by 253 words which are not in the dictionary. The "vocabulario commun" is more restricited. It consists of words found in all four sources listed under the usual vocabulary. The common vocabulary contains 1971 words and a supplement of words which do not appear in all four types of sources, but still have a total frequency of 40 or more, including some which reach 40 by combination of related forms of a basic word. The fundamental or basic vocabulary is the most restricted. The list consists of 208 words whose frequency is nearly equally distributed among all four sources, provided that the total frequency is above 40. Some 26 words of high frequency (over 400) were eliminated from the list because of unbalanced frequencies with respect to letters (as a source); 19 were too high and 27 were too low. In addition, there are sections on correlation among the four sources, on factorial anlaysis, and a conclusion.

113. Garvin, Paul (ed.) Natural Language and the Computer. New York: McGraw-Hill, 1963.

This is a collection of 16 original essays concerning all phases of computer-aided studies of language.

114. George, Alexander L. Quantitative and qualitative approaches to content analysis. Trends in Content Analysis, ed. Ithiel de Sola Pool, Urbana: University of Illinois Press, 1959, 7-32.

A survey of experimental design problems and methods of quatification.

115. George, H.V., An inventory of simple sentence patterns of English. Proceedings of the Linguistic Society of New Zealand, 1967-1968, 10-11, 62-66.

A brief inventory is presented here. The author demonstrates the possibility of a language model being constructed rather than presenting a comprehensive analysis of the English Language. This model, he asserts, would be of some value to teachers of English

115. (continued)

and might serve as a tool for future research. He describes how he constructed the model using 13 elements and trying to organize a system by hand. He found that to be of little value, and he used a computer which allowed him to list his elements and mutually excluding items. Then he requested permutations from the computer. These results were then culled for items that no examples could be constructed from. The results were then listed and produced a count of 518 patterns. The elements and codes were:

s.	Subject.
si.	Subject formal it.
st.	Subject formal there.
o.	Object.
p.	Predicative adjunct to the subject.
no.	Not.
ne.	Never.
f.	Finite verb, except am, are, is, was, were.
fb.	Finite verb, am, are, is, was, were.
a.	Auxilary except items of fb.
vs.	Non-finite verb stem.
vd.	Non-finite verb stem +ed.
vg.	Non-finite verb stem +ing.

His 12 most frequent patterns were:

o.
p.
fp.
fo.
a.
no.
p no.
no p.
no o.
no fp.
no fo.
ne a vs p.

His 12 least frequent patterns were:

ne f st p.
st ne a.
st ne a vd.
a st ne vd.
ne a st vd.
st a ne vd p.
p st ne a vd.
p st a ne vd.
a st ne vd p.
ne a st vd p.

115. (continued)

The main problem with the work is that it does not present a comprehensive analysis. The size of the corpus from which this count was drawn is not mentioned and the number of sources used is not discussed. The method of analysis is not explained deeply enough. However, the author's goal appears to be to establish a model for future research rather than a large analysis of the language or languages in general.

116. Gibson, James W., Gruner, Charles R., Kibler, Robert J., and Kelly, Francis J. A quantitative examination of differences and similarities in written and spoken messages. *Speech Monographs*, 1966, *33*, 444-451.

This study examined the possible differences and similarities in spoken and written style. Using 45 speech students, the authors had them write essays and make speeches on given topics. Using severals methods of analysis, they concluded that spoken style was more interesting and simpler to understand.

117. Gilmore, T., and Kwasa, S. *Swahili Phrase Book for Travelers*. New York: Frederick Ungar Company, 1963.

A little broader in scope than most word and phrase books for travelers, this book attempts to cover a broad area of dialectical variation with a single word and phrase list of "essentials".

118. Good, I. J. Distribution of word frequencies *Nature*, 1957, no. 4559, 595.

This is a very brief item on the relation between Zipf's rank-frequency hypothesis and Shannon's entropy.

119. Gougenheim, G., Michea, R., Rivenc, P., and Sauvage, A. *Elaboration on Fundamental French*. Paris, Franck: Didier, 1964.

The introduction explains the reasons for fundamental French in some detail. Part 1--Chapter 1 provides a history of simplified (basic) vocabularies. Chapter 2 discusses Basic English. Chapter 3 discusses statistical methods of language analysis indicating their history and a preference for the statistical methods over the logical/subjective methods. It refers to the

119. (continued)

French frequency dictionaries by Henmon (1924-400,000 written words) and discusses them briefly. Chapter 4 discusses scholarly words derived from the Vander Beke study, mainly Basic French Dictionaries. Chapter 5 deals with two lists, (1) the Aristizabal list of 1938 which was based on 1400 letters written by adults in several situations, 4100 compositions by children of various school grades and 25 stories invented and told by gifted children. The authors came up with 460,727 running words containing 12,038 different words of which 4329 had a frequency of greater than 10. The Dottreme-Massarents List which was based on prior lists (Aristizabel, Haygood, Vander Beke/Prescott) and on studies by Dottreme himself. The final vocabulary count contains 2750 words arranged by frequency, difficulty of spelling, and a quotient resulting from dividing the frequency by the difficulty of spelling. The Appendix to Part 1 refers to frequency counts in languages other than French and English. Part 2, Chapter 1 describes in detail the method of obtaining the samples of Fundamental French. It includes a list of 1063 words in order of decreasing frequency. Next is the list of alphabetical order. Chapter 2 contains studies on frequency/grammatical relationships. Chapter 3 discusses relationships between literary and spoken French and gives a table indicating Rank by Frequency and a value for the Zipf Constant (f x r) where a = 1.305 on a corpus of 312,135 words of which 7.995^a are different. Part 3 discusses the problem of availability versus frequency, the available vocabulary and the degree of its availability, the psychological stability of concrete words, sociological and geographical differences, and complementary research studies. Part 4 contains the vocabulary, including additions and deletions, notes on grammar, and Verification Measures. The appendices include Extracts of Recordings, Examples of Instructional Tests written in Fundamental French and a Bibliography.

120. Graham, E. *Basic English-Internation Second Language*. Orthological Institute, New York: Harcourt, Brace, and World, Inc., 1968.

This work combines and updates Ogden's *Basic English* and the *ABC's of Basic English*.

121. Green, J. R. A comparison of oral and written language: a quantitative analysis of the structure and vocabulary of the oral and written language of a group of college students. *Dissertation Abstracts*, (1959), 19, 2080-2081.

121. (continued)

> "From the language data, tabulations were made of the number of times each different word was used, the letter-lengths of each word, frequencies of parts of speech, frequencies of main and subordinate clauses, and frequencies of various kinds of subordinate clauses. Verbid equivalents of finite clauses were counted in a further study of ratio of subordination." The material studies consisted of 13,684 words of speech and 18,447 words of writing.

122. Greenway, P. J., A Swahili-Botanical-English Dictionary of Plant Names. Tanganyika (Tanzania): Dar-Es-Salaam, 1940. (2d Revised Edition).

> The book lists botanical names in both Swahili-English, and English-Swahili orders. Brief descriptions of each plant, tree, or shrub are provided as English translations for the Swahili terms.

123. Gross, M. Mathematical Models in Linguistics. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

> Structural linguistics deals with the properties of natural languages that are best accounted for in terms of combinations of simple elements into more complex ones. There are laws that restrict the combinations. In the last 20 years, research in linguistics has reached the degree of complexity and precision such that the use of mathematical tools has become the only safe way to state the descriptions. This book presents a number of such tools, in terms of standard mathematical notations. It also attempts, in a more general way, to demonstrate how the tools can be used in linguistics, especially in the construction of models.

124. Gruner, Charles R., Kibler, Robert J. and Gibson, James W. A quantitative analysis of selected characteristics of oral and written vocabularies. Journal of Communication, 1967 17 152-158.

> "The purposes of this study were: (1) to develop a list of the twenty-five most frequently used words for both oral and written messages; (2) to compare these word lists with similar lists developed from previous research; and (3) to determine the differences and similarities between written and spoken vocabularies as measured by the type-token ratio." Forty-five college students provided the data for the analysis.

125. Guilbert, Louis De l'utilisation de la statistique en lexicologie appliquee. *Etudes de Linguistique Appliquee*, (1963) *2*. 12-24.

This is a general consideration of the use of statistics in language studies with particular attention to the problems presented by idiomatic phrases, grammatical variants, and semantic relations.

126. Guiraud, Pierre *Bibliographie critique de la statistique linguistique.* Utrecht-Anvers: 1954.

A multilingual collection of books and articles arranged by content category.

127. Guiraud, Pierre *Les caracteres statistiques du vocabularie*. Paris: 1954.

More than half of the book is devoted to problems of the analysis of lexical distribution in literary texts. The concluding sections are devoted to a presentation of lexical data derived from the study of poems by Baudelaire, Rimbaud, Mallarme, Apollinaire, Valery, and Claudel.

128. Harwood, F. W., and Wright, A. M. Statistical study of English word formation. *Language*, 1956, *32*, 260-273.

"A quantitative study of English word formation based on the data of the Thorndike-Lorge frequency list. Results cover (a) dimensions of the word forming mechanisms in modern English, (b) measures of the relations between major suffixes and word classes, and (c) the main equivalencies symbolized by the major suffixes."

129. Haydon, Rebecca E. The relative frequency of phonemes in general-American English. *Word* (1950), *4*, 217-223.

This article presents the results of the analysis of six classroom lectures by different speakers transcribed in the system developed by Kenneth L. Pike.

130. Hays, D. G. *Introduction to computational linguistics*. New York: 1967.

> This is a textbook introduction with frequent illustrations of computer algorithms for linguistics and exercises for the student.

131. Henmon, V. A. C. *A French word book, based on a count of 400,000 running words*. (University of Wisconsin Bureau of Educational Research Bulletin No. 3, September 1924.) Madison, Wisconsin: University of Wisconsin College of Education, 1924.

> This is a count of printed and written French as of the time i.e., prior to 1924. The study details the particulars of its compilation. It was intended as a companion piece to Thorndike's "The Teacher's Word Book" on English and Kaeding's "Frequency Dictionary" ("Haeufigkeits-woerterbuch") on German. The 400,000 running words were reduced to 9187 on a dictionary basis. Of these, 3905 occurred five times or more. They are printed in the book in order of frequency (Part 1) and alphabetically (Part 2). Words occurring 5000 times or more account for 25% of running discourse. There are only ten such words, but they include the verbs "to be" and "to have" with all their conjugations subsumed under the infinitive. 655 words occurred 50 times or more and 1250 (including the 655) occurred 25 times or more. Unfortunately, Henmon did not give any details of the techniques involved in the corpus selection, he only indicated its general breakdown as to source. Nor did he indicate how he developed and refined the word count. However, the book is significant as one of the earlier word counts of some scope.

132. Herdan, G. A new derivation and interpretation of Yule's 'Characteristic' *K*. *Journal of Applied Mathematics and Physics (ZAMP)*, (1955), *4*, 332-34.

> "Yule's 'Characteristic' *K* of the word-frequency distribution of a linguistic text is derived under the assumption that the occurrence of a word in such texts was governed by a law of chance (the Poisson law). This assumption, and with it the use of *K* as a characteristic of the text, has

132. (continued)

> been attacked by linguists, without foundation, as the writer believes. However, the constant _K_ can be derived without such an assumption, which has not only the advantage of obviating adverse criticism of the kind referred to above, but of showing _K_ to be an easily interpretable, useful and interesting characteristic of a linguistic text."

133. Herdan, G. The relation between the functional burdening of phonemes and the frequency of occurrence. _Language and Speech_, (1958), _1_, 8-13.

> "The frequency of occurrence of phonemes in a language may be derived from dictionary material or from continuous texts. This paper deals with the relation between the two sets of values for English. When distributions are plotted for English phonemes, classified according to manner and place of articulation, it is seen that there is a close similarity between the distribution for dictionary material and for continuous texts. The hypothesis is advanced and tested that the phoneme distribution in speech is a random sample of the phoneme distribution in dictionary material (the functional burdening of phonemes)."

134. Herdan, G. _Quantitative Linguistics_. Washington, D. C.: Butterworths, 1964.

> The thesis of this book is that mathematical linguistics is an integral part of linguistics, and not just some tool used on an _ad hoc_ basis to obtain statistical data. Herdan's concept embraces and ties together deSaussure's and Bloomfield's and differentiates among:
>
> La Langue--the language viewed as an entity,
> La Langage--collective human speech as an entity (quantitative linguistics) which is somewhat different than La Langue, and
> La Parole--actual individual human speech or utterance which differs from both of the above.
>
> Herdan has divided his book into four parts and 20 chapters, together with an appendix which provides a numerical table of the law of solidarity (in the use of words it is the system of vocabulary as revealed in the gradient of frequencies). The four broad categories of analysis and exposition in Herdan's book are: (1) quantitative linguistics, (2) phonemic level, (3) vocabulary level, and (4) syntax level.

135. Herdan, G. The Advanced Theory of Language as Choice and Chance. Berlin and New York: 1966.

The most recent and most comprehensive of Herdan's textbooks, this work gives particular attention to matters related to stylistics.

136. Hibbett, H. and Gen. Itasaka Modern Japanese - A Basic Reader. Cambridge, Massachusetts: Harvard University Press, 1965, Volumes I and II.

Prepared with the assistance of an HEW (USOE) contract the two volume text contains vocabulary lists and notes (Volume I) and Japanese Text (Volume II). This is a textbook which should not be studied until after basic (or beginning) Japanese has been mastered. It attempts to use the most frequently used Japanese words as determined by the 1957 and 1960 vocabulary studies by the Japanese National Language Research Institute. Although it is in current or modern Japanese, it is written/printed word, rather than spoken word oriented.

137. Hill, Archibald Oral Approach to English. Tokyo, Japan: The English Language Education Council, Inc. 1965, 1 and 2.

This book consists of progressive drills in spoken English without reference to the origin of word selection. Drills: understanding sounds by contrast, producing sounds, use of sentence patterns and sequences, substitution frames, dialogues, and grammatical transformation practice.

138. Hill, L. Selected Articles on the Teaching of English as a Second Language. London: Oxford University Press, 1967 (1969 reprint).

The author has spent most of his adult life teaching English to foreigners in their own countries. He has also written many articles on how to do it. Eighteen of the ones he considers best he has inserted into this compilation. The articles are full of useful hints on how to teach English, in somewhat the same vein as Stevick's "Helping People to Learn English."

139. Holstein, A. P. A statistical analysis of Schizophrenic language: preliminaries to a study. Statistical Methods in Linguistics, 1965, 4, 10-14.

> This article contains statistical summaries and a brief discussion of "20 minute samples of the speech of 8 schizophrenic patients" with comparable data from published counts. The author calculates word class distribution, Yule's K, and the type-token ratio and lists the most frequently used words.

140. Horn, E. A Basic Writing Vocabulary. (University of Iowa Monograph in Education. First Series No. 4, April 1, 1926) Iowa City, Iowa: College of Education, University of Iowa, 1926.

> This is a vocabulary based on the 10,000 English words most commonly used in writing. It contains an identification and critical review of earlier writing vocabularies as well as the methods used in developing the 10,000 word vocabulary. The author points out the value of the list in teaching English to foreigners since the first 500 words common to his. Thorndike's lists, and spoken vocabulary lists make up 75-80% of the running words in English. This is a worthwhile, albeit somewhat dated study.

141. Horowitz, W. and Berkowitz, A. Structural advantage of the Mechanism of spoken expression as a factor in differences in spoken and written expression. Perceptual and Motor Skills, (1964), 19, 619-625.

> Type-token ratios were computed as part of this study of writing and speech.

142. Horowitz, M. W. and Newman, J. B. Spoken and written expression: an experimental analysis. Journal of Abnormal and Social Psychology, 1964, 68, 640-647.

> "Two experiments were designed to test for the differences between written and spoken expression. These two modes were controlled by limiting time for the preparation, time

142. (continued)

for exposition, and by limiting the subjects to two balanced topics. . . " Type-token ratios were calculated among other measures.

143. Howes, D. A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior*, 1966, *5*, (6), 572-606.

Howes undertook this research in order to up-date and correct what he considered deficiencies in prior counts, especially Thorndike's omission of spoken English; the French, Carter, and Coenig Telephone Count (1930) being designed to record speech sounds rather than words--a method of collection was not from running connected samples--sampling was restricted; Fairbanks (1944), small corpus--30,000 words-only those words with a frequency of 100 or more were published. Informants for the Howes' study were 20 sophomores at Northeastern University and MIT and 20 VA Hospital patients who had acted as controls for his prior studies on aphasic speech but were themselves free from cerebral defects or acute dehibilitating diseases. Informants were taped in free speech in response to general questions designed to get them talking naturally. All recordings took place between 1960 and 1965. There were 50 interviews of 5000 words each. The 41st (VA) informant provided 10 of the 50 interviews in order to provide data on stability of word frequency. The 40 others were each interviewed only once. The total corpus was 250,000 running words from 41 sources, which were catalogued as to individual source as well as to class of source; i.e., University or VA Hosptial. There were 9699 words in the corpus, of which a little less than half (47 percent) occurred only once. The author notes that the type/token ratio of spoken English tends to be less than it would be in written or printed English, and that only very large counts will produce evidence of extremely rare words. (Bongers says at least a million. This count is only 25 percent of that amount). The results are tabulated in an alphabetical list giving total frequency (all 41 informants) and separately University (20) and VA Hosptial (20) frequencies. The informant interviewed 10 times appears only in the total column. Words with a frequency of one are listed linearly to save space, but are annotated to indicated whether they were used by a University student, one of the 26 hosptial patients, or by the one VA patient interviewed 10 times. In spite of its limited sampling, this is a useful count since it is recent and embraced hospital patients from a variety of backgrounds (although probably mainly from the lower middle, and lower class) as well as students, most of whom were probably in their 14th year of the

143. (continued)

educational process. Within the student group, although Howes did not break them out as such, he had some variety, at least of academic interest, and probably also of socio-economic background.

144. Hultzen, I., Allen, H., Jr., and Miron, M. Tables of Transitional Frequencies of English Phonemes. Urbana, Illinois: University of Illinois Press, 1964.

The frequency of occurrence of transitional probabilities of small units in normal text may be different in sequences following any other given unit from what it is when the preceding unit is taken into consideration. A phoneme is defined as the least unit for which a distinction must be made in a language. Phonemic analysis yields more usable data than analysis by spelling letters. The objective is to set up an apparatus for describing the set of phoneme sequences occurring in a running text of language. The corpus used for the study was drawn from 11 different plays in the publication "Plays - The Drama Magazine for Young People," published by the Journal of Modern American English. Selections were one page each. They were run together to obtain a total of 20,032 phonemes, including junctures as phonemes. The phonemic analysis follows that Professor Agard used in the Southwest Project in Comparative Psycholinguistics. In this case, he spoke the selected excerpts of the 11 plays in his modified version of the southeast New England dialect. The phonemic notation was that used by Trager and Smith in their Outline of English Structure. In presenting the tables and corpus, an IBM printout was used with its limitation of capital letters and a few non-literal symbols. Tabular displays include the number of occurrences of single through four phoneme sequences. The fourth order sequences are also tabulated by reverse indexing. In Chapter 2, Section I, there are several breakouts and elaborations on the tables in Part II, including frequency by types and tokens. In Chapter 3, Section I, there is a discussion of messages generated by computer on the basis of transitional probabilities.

145. Ichiro, S. Basic Vocabulary for School Children (Kyoiku Kihon goi). Tokyo: Maki Shoten, 1958.

This vocabulary totals 22,500 words for use in the nine years of compulsory education of Japanese children. It is graded

145. (continued)

as follows:

lower Primary --- 5,000 words,
higher Primary -- 7,000 words, and
junior High ---- 10,000 words.

The words were selected on the basis of subjective criteria by a panel using dictionary sources.

146. Jakobovitz, L. A. Foreign Language Learning (A psycholinguistic analysis of the issue). Rowly, Massachusetts: Newbury House, 1970.

This book is an attempt to unscramble some of the confusion between language teaching (methods and materials) and language learning (psycholinguistics and human variables). It is has five chapters:

1. Psycholinguistic Implications of Teaching of Foreign Languages,
2. Psychological and Physiological Aspects of Foreign Language Teaching,
3. Compensating Foreign Language Instruction (Teacher/Learner/Researcher/Evaluator),
4. Problems of Assessing Language Proficiency (see items on testing), and
5. Foreign Language Aptitude and Attitude References (Bibliography).

147. Johnson, D. B. Computer Frequency Control of Vocabulary in Language Learning Materials. Instructional Science, Amsterdam, The Netherlands: Elsevier Publishing Company, March 1972, 1, (1), 121-131.

"Vocabulary is one of the major obstacles to attaining reading fluency in a second language...For efficient learning, the vocabulary systems must be structured in terms of frequency groupings so that the more frequent ones are mastered before the less frequent ones...The solution involves: (1) the establishment of various word frequency groups and (2) marking the word in the reading text so that the learner has a clear set of rational priorities. Statistical studies suggest that approximately 5000 most freuqent words constitute a minimum vocabulary for "liberated"

147. (continued)

reading and account for about 90 percent of the different words in an average text...the presentation of the higher frequency words within the 1000-5000 range should be sequenced by groups in terms of their relative frequencies. Each group might correspond to a particular level of language proficiency. This goal can be attained by means of a system in which the frequency category of each text word is marked so that the learner knows its relative importance and can structure his vocabulary acquisition accordingly. A marking procedure by frequency is integrated with a marginal translation or glossing routine. The article proposes a set of frequency groups and describes an algorithm for the implementation of a frequency identification and marking procedure on an IBM 360 copmuter..." Although the article is devoted to reading skills it has obvious application to oral vocabulary, once determined, and its integration into oral sentence patterns or other methods of learning conversation.

148. Johnson, F. A Standard Swahili-English Dictionary. (For the Interterritorial Language (Swahili) Committee), London: Oxford University Press, 1939.

Madan's Dictionary (1903) was based on the language of Zanzibar City. This update broadens the geographic base of the word coverage. Nouns and other forms derived from verbs are listed under the verb rather than separately. The dictionary includes loan-words from Persian, Hindi, Turkish, Arabic and neighboring Bantu languages, in addition to a small number of Portuguese, German and English borrowings. (See Berritt, D. V.'s Dictionary.)

149. Jones, L. V. and Wepman, J. M. A Spoken Word Count. (PHS Grant MH 01849 and M-10006 (University of North Carolina). PHS Grant MH 01876 (University of Chicago)). Chicago Illinois: Language Research Associates, 1966.

The vocabulary was compiled from English-speaking adults who were each asked separately to tell a story about 20 pictures in Murray's Thematic Apperception Test of 1943. It was discovered that in spoken language, 33 words account for 50% of the words used. (An analysis of the Lorge and Thorndike Word Lists shows 89 words are required to account for 50% of their written word sample.) The most frequently used

149. (continued)

words are used more frequently by speakers than writers. The book has three lists:

List A - 1102 words most often used by 54 speakers; each word has a frequency of at least 4/100,000,
List B - Words spoken by at least two of the respondents, arranged by grammatical class, alphabetically within class, and
List C - List B in completely alphabetical order, and including inflectional forms.

A table shows the ratio of male/female usage of words as well as the ratio of persons under/over 60 years of age, recognizing that education probably has more to do with the variance than the categories listed. Zipf states that there is a relationship between word length and frequency of use. This study supports his thesis up to words four letters in length; after that the relationship is not exact.

150. Jones, R. M. Situational vocabulary. _IRAL_, 1966, _4_, 165-173.

Relating to the concept of selecting vocabulary according to the idea of availability (disponsibility), Jones discusses objective means for selecting "centers of interest" by advancing fairly rigorous definitions of "situation" or "center of interest" and for using objective criteria to list the "centers of interest" which are to be investigated. Jones discusses "open" and "closed" situations, "positioned" and "unpositioned" situations and recommends the development of an "Aristotelian" hierarchy in classifying vocabularies by situation; in effect, a situational taxonomy.

151. Joos, Martin Review of Zipf's the psycho-biology of language. _Language_, 1936, _12_, 196-210.

A detailed critique of major significance. Joos proposes a modification of Zipf's rank-frequency equation.

152. Joos, M. _The English Verb Forms and Meaning_. Madison, Wisconsin: University of Wisconsin Press, 1968.

Joos says that if German is hard to learn because of its nouns, English is hard to learn because of its verbs. He divides his book into:

152. (continued)

Chapter I	Introduction,
Chapter II	Non-Finite Verbs,
Chapter III	The Finite Schema,
Chapter IV	Basic Meaning and Voice,
Chapter V	Aspect, Tense, and Phase,
Chapter VI	Assertion
Appendices.	

153. Jorden, E. The syntax of modern colloquial Japanese - *Language*, (*31*, No. 1 (Part 3), January-March 1955.) New York: Krauss Reprint Corporation, 1966.

The author states that her purpose is to give a systematic and complete description of the syntax of modern colloquial Japanese and incidentally to formulate a new technique for analyzing language. The study is based on a corpus of 60,000 spoken words from the Tokyo area. Most informants were men and women between the ages of 20 and 50, representing varied professions and family backgrounds. All were native speakers of Japanese, educated at least through high school level. Topics talked on were anecdotes, personal experiences, and conversations between individuals. Some spontaneous speech heard in Tokyo was also recorded. Some contemporary newspapers and magazine articles, some interviews, round-table discussions, dialogues, and comic strips, and some fiction were also added, so that in its entirely the study was not completely of the spoken language. However, the written material was recorded as spoken by an informant. Material of a formal written style was omitted. Utterances were broken down into successively smaller sequences until the maximally independent (IC) sequence was reached (Lexeme). All sequences were then categorized (classified). The dissertation describes its method, materials, procedures, the system of classification, and its application. This study contains two appendices:

Lexeme Classes and
Constituent Types (of sequences).

154. Josselyn, H. *The Russian Word Count and Frequency and Analysis of Grammatical Categories of Standard Literary Russian*. Detroit: Wayne University Press, 1953.

154. (continued)

This word count provides data dealing with the distribution of vocabulary and structural categories of standard literary Russian. The time-frame is the second quarter of the 19th century to the present (circa 1950). The time samples were taken as follows:

15% -- 19th Century,
25% -- 1900-1918, and
50% -- 1918 to about 1950.

The classification of samples according to style is as follows:

7% drama,
14% literary criticism,
20% journalism (wide scope within magazines and newspapers), and
59% fiction.

The material is condensed into six lists:

List 1 is 204 most frequently used words out of 150,000 running words,
Lists 2 through 5 are the first 2000 words in groups of 500, arranged in alphabetical order, and
List 6 is the next 3,000 most important words for 3d and later year students of Russian.

Tabulations do not include proper names with some exceptions, inflected nouns are entered only as the masculine singular, all inflected verbs are entered under the infinitive, dialectical items are entered separately except for verbs, and dialectical verbs are referred to in their proper infinitive.

There is a tabulation of grammatical usage of several well-known authors. Special computer source and punch cards were prepared for essential data of the categories desired.

The total number of running words examined was	1,000,000
The total number counted was	526,044
The different words recorded were	41,115
The total significant words published in the lists were	5,230

The final lists show both range, frequency, chronology (period), type literature, and conversational or non-conversational, source.

Lists are given in order of range rank. The index is alphabetical with a List Key for each word.

155. Kaeding, F. W. *Haeufigkeitswoerterbuch der Deutschen Sprache (Frequency Dictionary of the German Language)*. Berlin: Mittler and Sohn, 1898.

This book is in German. It is a frequency count of words and syllables in German and is one of the earliest still cited frequency counts. The book begins with a review of literature on the subject, requirements for such a study (especially for stenography), prior studies, lists of source materials, and procedures followed. Some 11,000,000 words and 20,000,000 syllables were counted. Their ratio in the study is then 1/1.83. There are several tables which present material alphabetically and in frequency rank order. In the main, alphabetical table inflections are listed under the headword. This was a comprehensive and thorough work for its time.

156. Karlgren, Hans Positional models and empty positions. in *Structures and Quanta: Three Essays on Linguistic Description*. Copenhagen and New York: 1963, 22-56.

A discussion of the value of statistical considerations in a slot-and-filler model of language.

157. Keil, Rolf-Dietrich, Einheitliche Methoden in der lixikometrie. *IRAL*, 1965, *3*, 95-122.

After discussing various problems associated with lexicology, the author proposes that the corpus used should contain at least ten million running words, with single text containing no less than ten thousand running words, and that the functional weight given to text classes should correspond to the relative importance of these classes in the language as a whole. An extensive bibliography accompanies this article.

158. Kihouka, T. *Japanese language guide for Secondary School Teachers*. South Orange, N.Y.: Seton University, 1964.

This guide has five parts: approach, planning, materials, when to use Hiragana, Katakana and Kanhii, and Evaluation.

159. Kochi, D. *Basic Japanese (Kiso Nippongo)*. Tokyo: Kokuseikan, 1933.

This is a Japanese parallel to Basic English by Ogden. The idea was to streamline Japanese for instruction and especially to aid in teaching Japanese to non-Japanese speakers. The book is divided into three parts: sentence rules, basic reader, and includes 1000 word vocabulary.

160. Kochi, D. *Shades of Japanese (Nippongo no sagata)*. Tokyo: Kaizosha, 1941.

This book is a group of articles by the author, including one called "Kisogo (Basic Japanese)". It also contains a basic word list of 1100 words.

161. Koutsoudas, Andreas M., and Machol, Robert E. *Frequency of occurrence of words: a study of Zipf's law with application to mechanical translation*. Ann Arbor: University of Michigan Engineering Research Institute, Report no. 2144-147-T, June 1957.

"Existing laws concerning the freqencies of words in language--specifically Zipf's and Joos' laws--are examined by means of new formulas which permit comparison of these laws with easily obtainable data. The laws are shown to be inaccurate and inadequate for predicting the size of dictionary necessary for mechanical translation, or the frequency with which words not in a dictionary of given size will be found. It is concluded that an empirical approach to this problem is most promising." Appendix A (pages 7-13) by George J. Minty summarizes the mathematical basis of the new formulas.

162. Kramsky, J. The frequency of articles in relation to style in English. *Prague studies in mathematical linguistics*, 1967, *2*, 89-95.

The investigation of the statistical distribution of definite, indefinite, and zero articles in contemporary English reveals that there are not significant differences in the usage of articles in various styles.

163. Kraus, Jirf K stylu soudobe ceske reklamy (on the syle of contemporary Czech advertising). *Nas Rec*, 1965, *48*, 193-198.

163. (continued)

A statistical comparison of broadcast advertising with that of newspapers, based on a part-of-speech count and a word repetition index.

164. Krishnamurthy, K. H. Psycholinguistic study of a schizophrene's speech.

Language and Speech, 1969, *12*, 256-257.

"An analysis of a schizophrene's speech using a phonological system of notation is presented here. Grouping the utterance data into phonemic and non-phonemic phonatory, the latter in turn into the normal and occasional etc., rather than phonemic and prosodic, is shown to be more comprehensive and useful. The system aims at incorporating many fresh utterance details like stretches, response time, rate of phoneme production, tone-accent distribution and the like in an edited transcript which is also serially numbered in such a way as to help pinpoint discussion of any portion. This is shown to be a useful method of bringing out many features of psycholinguistic interest, such as the general description of a subject's phonation for comparative study, the richness and close correlation of the devices to the mood and contents, etc. It also shows that the way of using phonatory devices in active speech is more varied than our native grammatical conceptions indicate and includes illustrations of semantic incoherence and a thought-type involved at many levels characterizing psychotic speech."

165. Kroeber, Karl *A computer analysis of fictional prose style*.

Washington, D.C.: Office of Education, 1966.

"Fundamental characteristics of fictional prose style were studied through systematic and objective analyses of novelistic syntax and vocabulary. Sample passages from the major novels of Jane Austen, the Bronte sisters, and George Eliot, as well as novels by 13 other authors were analyzed. Information on sentences, clauses, and words was coded and transfered to magnetic tape. Statistical tests were run on the data, and frequencies of syntactic patterns and vocabulary preferences were printed out. The primary conclusions of the study were (1) it is not possible to define the style of any novelist through simple statistical analysis of his grammar or his word choice, (2) novelistic style can be satisfactorily identified only in terms of multiple factors, many of which go beyond the level of syntax and vocabulary, and (3) further systematic study of fictional prose style should be based on automated analysis of tests, as the human analysis of texts requires an exorbitant amount of time."

166. Krohn R. English sentence structure (sentence patterns). Ann Arbor: English Language Institute, University of Michigan, University of Michigan Press, 1971.

As the title implies this book deals with patterns or frameworks rather than frequency counts.

167. Kublin, H. Useful Japanese pronounciation and basic words. New York: Japanese Society, 1961.

This booklet is very short and basic. It is for the traveler and beginning student. It has two sections: Alphabetical Lists of Words and Classified Vocabulary; e.g., Everyday Expressions and Date-Time Expressions. There is no indication of how the words and phrases were selected. It is essentially a short, traveler's word and phrase book.

168. Kucera, H., and Francis, W. Computational analysis of present day American English. Providence, Rhode Island: Brown University Press, 1967.

This analysis was performed by computer on a nearly 1 million word corpus of natural language text compiled in 1963-1964 at Brown University. It contains both lexical and statistical data. The purpose was to compile a corpus of printed American English rather than to develop a basic vocabulary of most common words. The corpus is divided into 500 word samples of about 2000 words each from continuous discourse. All texts were first printed in 1961 and represent a wide range of styles, i.e., 15 categories: press, 3 (reporting, editorial, review), religion, skills and hobbies, popular lore, literature and biography, miscellaneous government documents, learned and scientific, and fiction 6 (general, mystery, detective, science, adventure/western, romance, and love story). Samples were randomly selected. The analysis is in two main parts: word lists and statistical tables and graphs. Word lists are: descending order of frequency, alphabetical, first hundred most frequent words by total and the 15 categories, word frequency distribution, and sentence length distribution (corpus as a whole: 19.27 words; range: 25.49-12.76 for Government Documents (miscellaneous) and fiction/mystery, respectively.

169. Kucera, H., and Monroe, G. A comparative quantitative phonology of Russian, Czech, and German. New York: American Elsevier Publishing Company, Inc., 1968.

This is a book on computational linguistics financed, in part, by the National Science Foundation, Institutional Grants and Facilities Grants. The project reported on was designed to test the usefulness of well defined quantitative procedures in phonological analysis, especially comparative and typological studies, with emphasis on the latter. The study explores, in addition to phonotactics, the relative frequency of individual phonemes and phoneme strings, probabalistic constraints on the occurrence of phonemes in specified positions in relevant linguistic segments (i.e., syllables and words), or restrictions on sequences of larger phonological units. The research is valuable to historical phonology, revealing differences in historically related languages. The basic mathematical procedure uses the concepts of information theory. The first step was the phonemic analysis and transcription of a significant body of data in three languages. The corpus consisted of 100,000 phonemes for Russian and Czech and 105,174 for German. Sources for printed texts of 20th Century authors included 60 percent prose fiction, 20 percent journalistic press, 10 percent poetry, and 10 percent scientific and scholarly. The data were placed on punch cards in standard spelling with Russian transliterated into the Roman alphabet. An algorithm was constructed to transform the graphic presentations into a phonemic one. After a test, this part was done automatically. Some statistical counts were performed along with the transcriptions in Russian and Czech. The German text was pre-edited by separating prefixes from the item by using hyphens. German transcription was semi-automatic. After corrections, the statistical information was written onto magnetic tape. Chapter 4 is devoted to defining the phonological syllable. The three corpora were chosen to be comparable in content and style. Calculations were performed to determine entropy and redundancy. There is an Isotropy Index of two parts: Isotropy proper or phonotactics (matching phonemes in corresponding syllabic positions) and Isomorphy--quantitative similarity of phonemes. The concept of language divergence equals the difference between the actual Isotropy Index and the maximum possible value of the Index. This difference turned out to be least between Russian and Czech; middle for Czech and German, and the greatest for Russian and German (as might be expected). Conclusions: Close genetic relationships of two lanuages are likely to be shown at the phonological level in similar phonotactics, but not necessarily in very similar phonemic systems (as Russian and Czech). Languages in close contact (as Czech and German) may well show greatest similarity of phonemic inventory but less in phonotactical or phonological levels.

170. Lachman, R. Lachman word count frequency table. New York: Department of Psychology, State University of New York, October 1967 (a computer readout).

It is based on a corpus of 465,452 words, including punctuation. There are 18461 different words. In September 1965, 976 students of both sexes took 40 minutes each to write on any subject except psychology. There are two tables: alphabetical and frequency ranking.

171. Lado, R. Annotated bibliography for teachers of English as a foreign language. Bulletin 1955, No. 3, US Department of Health, Education and Welfare, USGPO 1955.

It contains material for the teacher, including tests and vocabularies or word lists and materials for students, with brief notes about each item.

172. Lado, R., and Fries, C. C. English sentence patterns. Ann Arbor: English Language Institute, University of Michigan Press, 1961, 1.

This book has for its purpose the understanding and production of English grammatical structure by means of an oral approach. It contains simple intermediate and advanced patterns. It is adaptable for use with various levels of student ability. It states that learning a new language consists not so much of learning about the language as in developing a new set of (thinking) habits. It has exercises for developing the new required new "habits". Each lesson has: an outline, a frame (including attention pointer, structural pattern, and comments), illustrative examples, practice exercises, notes, and a review.

173. Lado, R., and Fries, C. C. English pattern practice. Ann Arbor: English Language Institute, University of Michigan Press, 1958, 2.

Supplements Volume 1 with practice material. Procedures are entirely oral. The basis is a shift from mere imitation and repetition of patterns through conscious choice of elements of structure to be learned from exercises in which attention is centered upon a variety of lexical meanings substitutable in the structural frame. It is one of 34 units of the intensive course in English at Michigan. It is based on the idea that to learn a new language one must orally establish the patterns of the language as a subconscious habit. The pattern rather than particular sentences is the target of learning, i.e., the significant framework.

174. Lamb, Sydney M. The digital computer as an aid in linguistics. *Language*, 1961, *37*, 382-412.

A general introduction to computers and to problems solvable by techniques of "mechanolinguistics". (Available in the *Bobbs-Merrill Reprint Series in Language and Linguistics*, no. 56.

175. LeBreton, F. *Up-country Swahili exercises*. (for the soldier, settler, miner, and merchant and their wives.) Richmond, Surrey, England: R. W. Simpson and Company, Ltd., 1944.

This book tries to adapt the limited Swahili of the hinterland for use by individuals who have to move inland. It is largely a book of grammar, vocabulary and pronounciation. There is a special vocabulary on military terms, a Swahili-English and an English-Swahili vocabulary, and a key to the exercises.

176. Light, Richard L. *A study of some factors involved in teaching technical vocabulary to foreign military trainees learning English*. Master's Thesis in Applied Linguistics, Georgetown University, November 1964.

Ninety percent of the study group was foreign naval personnel (FY 64). The audiolingual approach to language was employed. Materials for teaching technical vocabulary in an aural-oral teaching situations were found to be lacking. The problems were: (1) finding an important technical field common to the majority of students (35 specialties involved), (2) criteria were word frequency, word importance, safety, and US Navy Word List for compiling the technical graded word list, (3) developing supplemental materials using pattern practice for word recognition and structure patterns, and (4) classroom trial of materials developed. Analysis indicated that electrical terms were the most common across the specialty fields. The criteria for construction of the word list were word frequency counts and the occurrence or non-occurrence of words in a US Navy List of electrical terms: Appendix A--Weighted Basic Terms in Electricity, Appendix B--Alphabetical List of Vocabulary, Appendix C--Lesson 1 (Patterns), Appendix D--Quiz with Pictures, and Bibliography.

177. Loogman, A. Swahili grammar and syntax. (Duquesne Studies--African Series No. 1) Pittsburgh, Pennsylvania: Duquesne University Press, 1965.

The author developed this text from his experience gained from 37 years in Swahili-speaking Africa and the teaching of the language. Part 1, Morphology, includes preliminary studies, nouns, qualifiers, substitutes (pronouns), verbs, adverbs, prepositions and conjunctions, idiophones, enonatapea, words, and interjections, and parsing. Part 2, Syntax, includes sentences, nouns, qualifiers, substitutes, verbs, binders, verb forms, auxiliaries, directive verbs, passive, to be, and to have. A bibliography is also included. Part 2 of this text is particulary valuable.

178. Loogman, A. Swahili readings. Pittsburgh, Pennsylvania: Duquesne University Press, 1967.

The purpose of this text was to help the student advance his study of the Swahili language from basic grammar and syntax to a profitable contact with well-written Swahili in order to provide an opportunity for observation, analysis and imitation. The materials were selected from a wide range of subjects and types and include: educational materials, histories, folklore, literary writing, journalistic material, oratorical material, letter writing, and poetry. The materials are divided into lessons each of which is accompanied by exercises in translating English into Swahili. A key to the exercises is at the end of the book.

179. Lorge, I., and Thorndike, E. A semantic count of English words, New York City: Institute of Educational Research, Teachers' College, 1938.

This is an account of the frequency of occurrence of each meaning of each word, i.e., a semantic count based on 2,250,000 words and the Thorndike 20,000 most common words (early version of the 30,000 word list). It is a hectograph reproduced in three-ring binders by alphabetical groupings.

180. Lorge, I. The semantic word count of the 570 commonest English words. New York City: Teachers' College, Columbia University, 1949.

This book contains the relative frequency of occurrence of the different meanings of each of the most common words. It supplements the Lorge and Thorndike List of 1938 for the 570 most common words in English.

181. Mackey, W. F., and J. G. Savard The indices of coverage: a new dimension in lexicometrics. IRAL, 1967, 5, 71-121.

Describes research into the development of indices of coverage or availability. The usefulness of a word considers the power of a word to define, to extend its meaning, and to include or to combine with other words. The authors conclude with a table of 3,626 words arranged in decreasing order of index of coverage, together with separate ratings for definatory combinational, inclusional, and extensional power.

182. Mackey, W. F., Savard, J. G., and Ardouin, P. Le Vocabulaire Disponible du Français (The vocabulary of available words of the French language). (In French) Montreal: Didier, 1971, 1 & 2.

The purpose of Volume 1 is to document the differences and similarities of concrete words used in France and in Acadia. The purpose of Volume 2 is to concentrate in more detail on the concrete words as used in Acadia, documenting the findings on Acadian children according to age and considering the effects of bilingualism. For Volume 1, word usage was tabulated in New Brunswick, Canada and four regions of France. The sessions with the informants were held from 1961-63, with the majority in 1962. In Canada, the sessions centered around 22 areas of interest (27 for bilingual children). The informants were 1745 school children from ages 9-18, located in 47 classrooms in 19 schools scattered throughout New Brunswick. The total corpus numbered 900,000 words. Concrete vocabulary was elicited by using as stimulus words the basic word of the center of interest, such as "animal", "body", and "transportation". Each child was given 15 minutes to write all the words he knew related to the center of interest. Only 2-3 centers of interest were covered at each session. In France, the informants were about 700 school children ages 9-12 in about 20 classes in as many schools. the total corpus numbered 300,000 concrete words as derived from over 16 centers of interest. The indices used to determine vocabulary were: frequency of use, distribution (number of persons writing each word), valence (the powers of a word to combine into compounds or idioms, to act as a synonym for another, to explain other words, and to express completely or slightly different meanings). The running words in the corpera were reduced down to 10,000 different words but these 10,000 were expressed in some 64,000 forms. (i.e., the average word was spelled in six different ways by the 2500 children involved). On the average, the 10,000 words were used by at least 27 children. However, closer analysis revealed that almost 5,000 were used by only one child. That means that the common vocabulary is 5,000 concrete words or less based on their use by more than one person.

183. Malcolm, J. *A classification of five thousand words most commonly used in writing, as compiled by Dr. Ernest Horn in accordance with the principles of the new standard course--Pitman shorthand.* Masters' Thesis, New York: Teachers' College, Columbia University, 1939.

> This is a comparison of the Manual on "New Standard Course--Pitman Shorthand with "A Basic Writing Vocabulary of 10,000 Words by Dr. Ernest Horn. Ms. Malcolm's analysis indicates that the Pitman Manual required revision to make it conform to actual word frequency usage as reflected in the Horn list.

184. Mandelbrot, Benoit An informational theory of the statistical structure of language. In *Communication theory*, ed. Willis Jackson, New York and London: 1953, 486-502.

> The author concludes his discussion of statistical models and Saussurean linguistics with the observation that "a quite general statistical structure, entirely independent of meaning, appears, underlying meaningful written languages."

185. Marchand, H. *The categories and types of present-day English. Word formation (a synchronic-diachronic approach).* Wiesbaden, West Germany: Otto Harressewitz, 1960. (Also Auburn, Alabama: Alabama Linguistics and Philosophical Series No. 13, University of Alabama Press, 1967.)

> Although the author calls his approach synchronic-diachronic, he starts off by emphasizing it is meant to be up-to-date, although not all inclusive, preferring general types of words to their variations. Historical data on word changes is used only incidentally. After the introduction, the author deals with compounds, prefixation, and suffixation. He then proceeds in less detail to cover zero-morphemes, back-derivation, phonetic symbolism, ablaut and rime combinations, clipping (omitting part of the word in speaking), blending, and word manufacture. This book presents a comprehensive picture of the composition of English words.

186. Marchand, M. *Five thousand Frenc idioms.* Paris: Em Terquem, 1910.

> This is a book for advanced students who are learning French as other than their native tongue. It includes Gallicisms, proverbs, and idiomatic adverbs, adjectives, and comparisons. This edition

186. (continued)

> is a revision of an earlier work encompassing some 4,000 idioms. The scope of the book embraces some 170 subject areas giving words and expressions valuable to enlarging vocabularies of those who already know some French. Unfortunately, the author does not explain his sources very well and the book is not current by some 60 years.

187. Martin, S. E. *Basic Japanese conversation dictionary*. (Revised and Enlarged) (English-Japanese and Japanese-English) Tokyo: Charles E. Tuttle Co., 1963 (8th printing).

> This Dictionary contains 3,000 "useful" English words with their most frequent meanings and their Japanese equivalents. It is meant for use with Martin's works on easy Japanese and essential Japanese. Unfortunately, it gives no rationale for the selection of the words it contain.

188. Martin, S. Morphophonemes of standard colloquial Japanese. *Language*, New York: Krauss Reprint Corporation, 1966 (originally July-September 1952) *28*, (3) (Part 2).

> This study represents the first attempt to make a systematic study of Japanese morphophonemes on a synchronic level. An attempt was made to keep the analysis on a formal level, separate and distinct from semantic correlations.

189. Martin, S. E. *Easy Japanese--a direct approach to immediate conversation*. (3rd revised edition) Tokyo: Charles E. Tuttle Co., 1968 (17 printing).

> This book has four parts: Say it with a Word or Two (Lessons 1-13), Add a Bit of Action (Lessons 14-20), Sprinkle in a Few Particles (Lessons 21-30), and 3000 Useful Japanese Words. The Japanese-English part of Martin's Basic Japanese Conversation Dictionary.

190. Maw, J. *Sentences in Swahili--a study of their internal relationships*. London: Luzac and Company, Ltd., 1965.

190. (continued)

This study was originally a Ph.D. Dissertation for the University of London. The theory and method are those of Professor M.A.K. Halliday. The materials were collected in 1964-1965 near Tanga, Tanzania. They were almost entirely spoken and spontaneous. They deal with units larger than the word from a syntactic rather than from a morphologic point of view. The book does not carry the study down into the structure of the word and morpheme. The study material was taped in the form of conversations between native speakers, stories, anecdotes, and discussions. Some expert testimony of scholars was added to the field research.

191. Maw, J. Review of 'Swahili readings' by A. Loogman. Journal of African Languages, Hertford, England, 1968, 7 (Part 1).

Maw says that 'Swahili readings' is a collection of Swahili texts from various sources and exemplifying different styles of Swahili writings; some by natives, some not. Unfortunately, the author has not indicated the source well enough to permit knowing which author is a native speaker. The book was intended to help students improve their Swahili. It failed in its purpose because Father Loogman did not take the time to analyze the texts and derive useful lessons and experience from them.

192. Mayaji, Hiroshi. A frequency dictionary of Japenese words. Dissertation Abstracts, 1967, 27, 34424-43A.

The dictionary is the result of a count of 250,000 words from five writing types: fiction, drama, didactic prose, periodical writing, and scientific writing.

193. McCalla, Gordon I. and Sampson, Jeffrey R. MUSE: A model to understand simple English. Communications of the ACM, 1972, 15(1), 29-40.

"MUSE is a computer model for natural language processing, based on a semantic memory network like that of Quillian's TLC. MUSE, from a Model to Understand Simple English, processes English sentences of unrestricted content but somewhat restricted format. The model first applies syntactic analysis to eliminate some interpretations and then employs a simplified semantic intersection procedure to find a valid interpretation of the input. While the sematic processing is similar to TLC's, the syntactic component includes the early use of parse trees and special purpose rules. The "relational triple" notation used during

193. (continued)

interpretation of input is compatible with MUSE's memory structures, allowing direct verification of familiar concepts and the addition of new ones. MUSE also has a repertoire of actions, which range from editing and reporting the contents of its own memory to an indirect form of question answering. Examples are presented to demonstrate how the model interprets text, resolves ambiguites, adds information to memory, generalized from examples, and performs various actions."

194. McCarus, Ernest N. and Rammuny, Raji M., Word count of elementary modern literary Arabic textbooks. University of Michigan, 1968.

"A computerized word count is presented of 11 elementary Modern Literary Arabic textbooks used in the United States. The word count was started in 1967 to provide a practical vocabulary base for a fully-programmed self-instructional course on the phonology and script of Modern Literary Arabic. The first part of the count is a cumulative list (complied with the aid of an IBM 360/20 computer and an IBM card sorter) arranged alphabetically by Arabic root, according to conventional dictionary practice, of all the words listed in the 11 Arabic texts, with their English meanings, the sources for each word are given. The number of different textbooks in which the word occurs is indicated, as well as its frequency in the Landau word count (1959). Plurals are listed separately but following the singulars, and the imperfect tense of the verb is likewise listed following the perfect, if both occur. Homonyms having distinct plurals are listed as separate items. The second part of the word count consists of alphabetical list of the words occurring in all 11 textbooks, in 10 of them, in 9, and so on. A list of the 11 textbooks covered in the count is also included.

195. McGovern, W. Colloquial Japanese. London: Routledge and Kegan-Paul, Ltd., 1968.

This is essentially a Japanese grammar based on British experience in training naval personnel in intensive Japanese classes. The student gets a general survey of the scope of the language before being introduced to the details with graded exercises. It contains a Japanese-English vocabulary.

196. Meier, Helmut. *Deutsch sprachstatistik*. Hildesheim: 1964.

The book provides numerous data on German (esp. phoneme and word statistics) based on Kaeding's frequency dictionary and on a variety of continuous texts representing different styles.

197. Milic, Louis T. *Style and stylistics: an analytical bibliography*. New York and London: 1967.

Some eight hundred items devoted to stylistics arranged chronologically in five sections: Theoretical, Methodological, Applied, Bibliographies, and Omnibus Works. Items are annotated and indexed subject and topic.

198. Miller, G. A. *Language and communication*. (revised edition) New York: 1963.

Chapter 4, "The Statistical Approach" (pages 80-99) gives a survey of major studies in statistical linguistics and introduces the student to the basic problems of the field.

199. Miller, G. A., Newman, E. B., and Friedman, E. A. Length-frequency Statistics for written English. *Information and Control*, 1958, *1*, 370-389.

"The results of a tabulation of word frequencies in a sample of written English are analyzed in terms of word length and syntactic function. It is found that a simple stochastic model gives a rough prediction for the results obtained when all words are combined, but not when words are classified as function or content words. Function words are short and their frequency of occurrence is a decreasing function of their length; content words are longer and their probability is relatively independent of length." Zipf's and Mandelbrot's "laws" are discussed.

200. Moore, W., and Ogawa, Y. *400 sentence patterns with creative sentence patterns (English)*. Japan: Hosei University Press, 1954.

This text designed for use of Japanese students of English has three parts: Central Problems of Grammar--40 fundamental or elementary English sentence patterns. For each pattern there is a creative sentence pattern designed to stimulate use of the basic pattern. There is a vocabulary of families of words for varying

200. (continued)

the word use in the sentence patterns. Also included are pronouns, the verbs "to be" and "to have"; auxiliary verbs, and the present and present progressive tenses; Intermediate Patterns--additional tenses of verbs, conjunctions, idiomatic expressions, and additional families of words; and Advanced Patterns--additional conjunctions, gerunds, quotations, relatives, infinitives, and subjunctives.

201. Morgan, B.Q. German Frequency Work Book. (American and Canadian Committees on Modern Languages) New York: The Macmillian Company, 1931, 9.

This study revises Kaeding's Haeufigkeitswoerterbuch der Deutschen Sprache (see above) and uses its findings in the construction of a German vocabulary for teaching purposes. To correct Kaeding's work, Morgan reduced the words to their stems and he describes the system he used to accomplish that. The author admits that there are limitations to the study in terms of its age (1898), but he feels that Kaeding's wide use of sources and his large number of running words justify its use. The author also presents two word lists which were the results of his study. The first list shows the basic words he derived by using stem words and those which had a frequency of 200 or more. The second list is an alphabetic list of the words with their frequencies.

202. Muller, Charles Le MOT, unite de texte et unite de lexique en statistique lexicologique. Travaux de linguistique et de litterature, 1963, 1, 155-173.

A detailed discussion of the problems of defining the "word" for lexicographical and statistical purposes.

203. Muller, Charles Frequence, dispersion et usage: a propos des dictionnaires de frequence. Cahiers de Lexicologie, 1965, 7, 33-42.

This article argues that both frequency and dispersion must be taken into account in the preparation of word lists for language teaching. The Frequency Dictionary of Spanish Words (Juilland and Rodriguez) is discussed in some detail from this point of view.

204. Muller, Charles, Fréquence des signifiés ou fréquence des signifiants. Etudes de Linguistique Appliquée, (In French) 1971, 2, 74-87.

204. (continued)

> "A comparison of the frequencies of French words with those of Spanish words having the same semantic contents, according to "The Romance languages and their structures" by S. Juilland. Passing from the total of frequencies to the fundamental data provided by each word, a correlation is revealed between the frequencies observed in the two languages. It appears obvious that the frequency of words depends on the stylistic situation represented by the categories of texts used for each group of words. The probability of using a lexical element is determined much more by the situation than by the lexical structure of the language, and frequency is related to the signifiant as much as to the signifié."

205. National Institute of Health Seminar on computational linguistics. (Public Health Service Publication #1716) Washington, D.C.: Department of HEW, October 1966.

> This is a report of a seminar among linguists and National Health Service Personnel. There were 13 presentations. Most deal with machine analysis of language with emphasis on syntax, primarily and semantic meaning, secondarily. It is a valuable document in revealing trends in linguistics, particularly that of syntax versus phonology, the increased attention being given to semantics, and the use of computer assistance in langauge studies.

206. The National Language Research Institute (of Japan) A research of newspaper vocabulary. Tokyo: Yoyuya, Sinzyuku, 1952 (In Japanese).

> The main parts of this report are outline and scope, lists of words used in a month in a newspaper including words used more than 10 times, words used more than 100 times listed in order of frequency, and analysis (frequency of words by day, frequency by article, news item and classification by parts of speech). These words are listed in Japanese Alphabetical order. Although useful this research suffers from being more in depth than breadth, i.e., only one newspaper was used. It is, however, reasonably current, having been conducted in 1951-1952.

207. The National Language Research Institute. Research in modern vocabulary (Gendaigo no goi chosa) Tokyo: The National Language Research Institute (kokuritzu Kokugo Kenkyugo), [Part 1 (1953) & Part 2 (1958)].

Part 1, Research on Vocabulary in Women's Magazines (Jujin Rasshi no Yogo), was based on sampling the text of one year's issues of two representative women's magazines (3 million running words). Part 2, Research on Vocabulary in Cultural Reviews (Sogo Zasshi no Yogo), was based on a sampling of 13 cultural reviews (230,000 running words). About 4000 most frequently used words are listed in each case. The analyses consider mainly the statistical and semantic structures of the vocabulary and word construction. Procedures used are spelled out in detail. Much use is made of statistical sampling, as opposed to word count methods as used by Thorndike and Horn.

208. The National Language Research Institute (of Japan) Research on the vocabulary in a newspaper in the early years of the Meiji Period (1877-1878). Tokyo: Kanda-Hitotubashi Tiyoda, 1959.

There are five main parts to this study: an outline (and scope), procedures, tables (vocabulary; high and low frequency words, supplemental words, prefixes, and suffixes), analysis (symbol combinations; words of three Chinese characters style and vocabulary), and an appendix, including technical terms used. This is an interesting study although obviously dated.

209. The National Language Research Institute (of Japan) The use of written forms in Japanese cultural reviews. Tokyo: Kanda-Hitotubushi, Tiyoda, 1960.

This report of research contains two main parts: an outline (and scope) and lists and tables including list of words with two or more variants, table of frequency distribution of Chinese characters, frequency tables of Chinese characters with frequencies of one or more with their different meanings. At the end of this list is a supplemental list of 1850 Chinese characters in official use in Japan, and frequency of Chinese characters not in the official list, and a list of such characters.

210. The National Language Research Institute (of Japan) Vocabulary and Chinese characters in ninety magazines of today. Tokyo: National Language Research Institute, 1962.

> This study is in three volumes. Volume 1 is a general description of the project and vocabulary frequency tables. The samples are dated 1956. Fields covered include culture, business, popular science, housekeeping, sports, and other amusements. The sample used contained 540,000 words from a possible 140 million. After an introduction, the analysis is tabulated in a series of tables: 7200 Most Frequent Words in Alphabetical Order with Their Relative Frequencies, 7200 Most Frequent Words Arranged in Order of Frequency, Frequency Tables in Five Strata by Class of Magazine from which subsamples were taken, Bound-Form Frequency Tables. An appendix is included giving a justification and procedures followed. Volume 2, Chinese Character Frequency Tables, 1963, after an introduction, consists of a series of tables: Most Frequently Used 1995 Chinese Characters According to Relative Frequencies, Most Frequent 1995 Chinese Characters with their Different Meanings and Uses, and 3328 Chinese Characters used in Japanese arranged in their (Japanese) Alphabetical Order. Volume 3, Analysis of Results, 1964, is arranged under the following headings: Tables contain the 1200 most frequent fundamental words and the semantic classification of the 700 most fundamental words, statistical structure of vocabulary, usage of bound forms with frequency tables, means, and uses as pause groups or markers, an analysis of 4,381 compound words, discussion of formally similar words as different or same words, using a 974 word list and two approaches. This is a current analysis of part of the printed Japanese language.

211. Newman, Edwin B., and Waugh, Nancy C. The redundancy of texts in three languages. Information and Control, 1969, 3, 141-153.

> "The procedure that predicts the mean information per letter in a long text by adding the constraint measured between pairs of letters in a text has been tested more fully. Results are presented to show that with randomized texts there is a close approximation to the Miller-Madow prediction of simple bias. Their samples of English of varying complexity show slightly more information per single letter and much more information in an average letter for the more difficult material. Conversely, samples for Samoan, English, and

211. (continued)

Russian show some constancy in the average information per letter in spite of wide differences in the size of their alphabets. Thus, greater redundance is correlated with a larger alphabet. The three samples of English considered are from the Bible, William James, and the Atlantic Monthly.

212. Nice, Margaret Morse On the size of vocabularies. American Speech, 1926, 2, 1-7.

This is a general consideration of the problem of determining the extent of an individual's vocabulary.

213. Nisbet, J. D. Frequency counts and their uses. Educational Research, 1960, 3, 51-64.

The author focuses on the history of vocabulary counts and concludes that their value may not be as great as is usually supposed.

214. Oettinger, Anthony G. Linguistics and mathematics. Studies presented to Joshua Whatmough, ed. Ernest Pulgram, 's-Gravenhage: 1957, 179-186.

A discussion of the notion of "model" in mathematical linguistics with particular attention to those proposed by Condon, Zipf, and Mandelbrot.

215. Ogden, C. The general basic English dictionary. London: Evans Brothers, Ltd., 1960.

This volume uses the 850 basic words and 50 additional international words to explain 40,000 meanings of 20,000 English words.

216. Palmer, H. A grammar of English words. London: Longmans, Green, and Scott, Ltd., 1938 (1967 edition).

This book would be more properly entitled "A Grammatical Dictionary of English Words". The author gives 10,000 English words with their pronounciation, information on several meanings; the inflections, and derivatives, and the context

~~261~~ 262

216. (continued)

in which the word appears. (Collocation and phrases.) There are four appendices: verb patterns, important grammatical categories, measures of time, and irregular inflections.

217. Palmer, H. E., and Hornby, A. S. Thousand word English. London: George G. Harrap and Company, Ltd., 1937.

The authors divide this short book into parts: an introduction in which they discuss how the words were selected [a combination of methods: subjective; objective (quantitative) and empirical], and the vocabulary itself, including inflected forms (which raise the total real-world words to well over 1000). This is an interesting work on vocabulary selection by the use of prior studies revised in the light of experience and personal judgment.

218. Perrott, D. V. Concise Swahili and English Dictionary. London: English Universities Press (EUP), 1970.

This dictionary starts off with a concise grammar from her "Teach Yourself Swahili Book". It is followed by two sections: Swahili-English and English-Swahili; both with notes. The dictionary contains all the words heard by the author during her 30 years in East Africa, plus a selection of words from Krapf, Sacleux, and Madan. The load-words given are mostly from Arabic and Hindi, some from the Portuguese and German, and a large number from English. In this latter respect it differs from the Johnson Dictionary.

219. Petty, W. T., Herold, C. D., and Stoll, E. The state of knowledge about teaching vocabulary. (Cooperative Research Project No. 3128, Contract OE 6-10-120) Champaign, Illinois: National Council of Teachers US Office of Education, 1968.

The focus of the project is on the teaching of vocabulary rather than on developing it. It is also pitched towards native speakers of English. types of vocabulary are form (words or phrases) and type (speaking, listening, reading, or writing). Other subdivisions are formal, informal, or colloquial. It advises that the teacher decide on the

219. (continued)

vocabulary to teach then on the aspects of language such as grammar, phonology, semantics, and situations of verbal contexts. Chapter 5 discusses research design for vocabulary studies; a type vocabulary; functions of vocabulary, and a sample population.

220. Pfeffer, A. *Index of English equivalents for the basic (spoken) German word list*. (Grundstufe - 1st Stage) Englewood Cliffs, N. J.: Prentice-Hall, 1964.

This book contains the English equivalents of the meanings of the basic (spoken) German. The procedure Pfeffer used paralleled that of Lorge and Thorndike in prorating the relative frequency of a particular meaning to the frequency of occurrence of the word. Computer assistance was used where appropriate. The corpus of the semantic count was derived from taped interviews. Both frequency and range were listed in the semantic frequency count as well as in the original frequency count. (Range is the number of speakers who used the word as compared to the total number of speakers contributing to the sample.) There were some shifts in words from the basic count because of semantic importance and 355 subsidiary word forms were added (16 nouns, 76 verbs, and 193 adjectives, 16 adverbs, 44 pronouns and 10 contractions). The study of semantic meaning helped discover many synonyms resulting from the spectrum and diffusion of meaning of each word. 1277 words were listed finally. Only the meanings of greatest importance as indicated by actual usage were included in the list. The student learned load aided by semantics is indicated by the fact that the basic list of nearly 1500 words become 25,000 when major meanings were considered.

221. Pfeffer, A. *Basic (spoken) German Idiom list*. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

This is the third in the Pfeffer series of studies on basic spoken German. Idioms are restricted word patterns which are the substance of communications. They range from word pairs to whole sentences. The meanings of some are self-evident. The meaning of others is not. All are characterized by some form of interdependence of parts and have some meaning different from their parts

221. (continued)

taken separately. Idioms may be grouped as stylistic (proverbs and common places), linguistic (the degree of restriction of word colocation), and syntactic (grammatical combinations or formulas). Pfeffer defines what he means by an idiom and discusses prior idiomatic lists based on 19th and 20th Century printed prose such as Kenniston (Spanish), Cheydleur (French) and Hauch (German). The three above mentioned lists were subjective analyses. This Pfeffer list is based directly on basic spoken German machine-counts using 595,000 punched and coded cards which enabled the determination of unrestricted words and those in groups of restricted patterns. This list also uses some phrases of the utility and empirical words in the basic spoken list. Some 7500 oral patterns were identified. These were reduced to 1026. An additional 99 (out of some 1800) derived from spontaneous adult writing relative to the utility and empirical words in the basic list were added. The 1026 idioms were restricted in usage to an average of 15 percent of the time. However, 1125 (1026 and 99) represent about 85 percent of German oral idiomatic usage. Interestingly, the percentage of the words in the idioms is greater than 15 percent of the basic vocabulary. Also, the percentage frequency of the idioms is high. The idioms listed have a frequency/range (f/r) index of 3/2 or greater. Idioms are recorded generally in groups of mutual key words and arranged alphabetically. they also contain cross-references to the component words. Such a list as this is indispensable to teaching German. With the other two lists, there are some 6,000 meanings and expressions which, if learned, will make a student conversant with 85 percent of oral German.

222. Pfeffer, J. Basic (spoken) German word list. [HEW (Office of Education) Contract SAE 8824 and OE2-14-036] Englewood Cliffs, N.Y.: Prentice-Hall, 1964.

In this work, Dr. Pfefer was in close touch with the developers of basic German (Advisory Research Council of the Institute of Basic German) and the authors (Goughenheim and Rivenc) of Fundamental French. In his introduction, Dr. Pfeffer contrasts subjective and empirical approaches to word counts. He finds that the objective counts for use in reading vocabularies gave way in the 1950's to the use of the phonograph and tape recorder to record and analyze spoken language. He also includes an excellent resume of prior works in the field of vocabulary counts. Pfeffer uses what he considered to be the best aspects of word collation of the Spanish word count prepared by the University of Puerto Rico and of Funda-

222. (continued)

mental French. He used 595,000 running words based on 40 tape recordings made in Germany, Austria, and Switzerland. Each tape ran for 12 minutes. In addition, he obtained "utility" words from 21 tapes on which he collected material from 2000 pupils. These tapes yielded 420,000 verbs and adjectives as well as 420,000 nouns. Tapes were transcribed and each spoken word and word form, adequately coded, was transferred single and in context on a separate punch card. Excluded were proper names, place names, and adjectives derived from them, as well as hesitations, repetitions, and abandoned starts. The 595,000 cards yielded 25,000 lexical units. Range and frequency were computed. The frequency used indicates the sum of the frequencies of inflectional forms. The 1000 most common words (i.e., those with a frequency of 40 or more and a range of 25 or greater) were reduced by criteria of applicability, universality, and indispensability to 737 spoken words. Topical or utility counts were made in 82 intermediate and high schools in 48 cities in Germany, Switzerland, and Austria. This was done by association with 20 nouns, 12 verbs, and eight adjectives in a ten-minute period. This yielded 833,000 terms including 19,700 nouns, 7,400 adjectives, and 6,000 verbs. Of these, 347 were finally selected for inclusion in the list. Emphasis was placed on applicability as opposed to topicality which resulted in one-third of the words selected having an order of rank of 200 or below. The 737 words were combined with the 347 and rechecked for topicality limitations, and then were augmented by 185 carefully selected words based on direct or association sequences; words linking the specific to the whole and vice versa, missing opposites, basic derivatives, topical gaps (e.g., months, metals) and notions such as "deaf". The total count numbered 1269. It is arranged first in alphabetical order indicating families, second by parts of speech, and third in order of frequency and origin.

223. Pfeffer, J. Alan Grunddeutsch, Basic (spoken) German word list, Mittelstufe, Pittsburgh University Institute for Basic German, 1970.

"As a link between the words in everyday use and the sophisticated language of the arts and sciences, the 1,536 words of the "Mittelstufe" or Level 2 derive in nearly equal proportion from three sources: (1) the spoken or topical language, (2) a collation of all significant word lists compiled prior to February 1965, and (3) a statistical analysis of some 500,000 words in context published or reprinted during the years immediately preceding. The purpose of the list is to provide the lexical basis for teaching

223. (continued)

German in the third and fourth year in high school or the second year in college. Alphabetized word lists and appendixes indicating frequency of usage are included. Extensive reference to source materials is made according to topical listing."

224. Pimsleur, Paul Semantic frequency counts. _Mechanical Translation_, 1957, _4_, 11-13.

A consideration of the problems involved in making semantic counts.

225. Plath, Warren Mathematical linguistics. _Trends in European and American Linguistics, 1930-1960_, eds. Christine Mohrmann, Alf Sommerfelt, and Joshua Whatmough, Utrecht and Antwerp: 1961, 21-57.

A survey of the field with an extensive bibliography; see especially "Statistics of Sytle and Authorship", pages 27-30.

226. Polome, E. C. _Swahili language handbook_. Washington, D.C.: Center for Applied Linguistics, 1967.

This book covers a lot of information on Swahili. It presents the phonetics and morphology of Swahili systematically in modern terms. The section on phonetics was the most advanced to date in 1967 (see review by Maw). It begins with an introduction which covers the historical and geographical aspects of the language, then goes on to sketch its structure, written language, contrasts with English, and literature. The language used is that of a cultivated speaker of Zanzibar and of the Mrima coast.

227. Polome, E. C. Lumbumbashi OSwahili. _Journal of African Languages_, Hertford, England: 1968, _7_, (Part 1) 14-25.

This article focuses on the characteristics of the creolized variety of Swahili spoken by individuals in Lubumbashi (Elizabethville, Katanga Province, Republic of Zaire). Zaire Swahili is a distinct variety, in any event, but this article fixes on individuals with no formal education in East Coast (Standard) Swahili and who are residents of Lubumbashi. Lubumbashi Swahili is most like the Zaire dialect of Swahili called Kingwana. It contains many French loan words, some of which have changed meaning as

~~266~~

267

227. (continued)

used locally. To a lesser extent, English words have also been imported with workers from South Africa. There has also been an influence of local native languages including spelling changes as well as phonetic shifts. Morphological changes have occurred and possessives have been simplified. Syntax of Katangan Swahili has not diverged much from East Coast patterns (similar to Central Bantu to which it is related), although some French patterns have been superimposed on the original native patterns. In the lower classes, some of the words of East Coast Swahili have been lost and remaining ones have been forced to take on multiple meanings to maintain flexibility of expression. Changes are so great that colloquial uneducated speech in Katangan Swahili would not be understood on the East Coast, although that of the better educated classes in Katanga would be understood on the coast, albeit with some difficulty.

228. Posner, Rebecca The use and abuse of stylistic statistics. Archivum Linguisticum, 1963, 111-139.

A critical survey of the field with comments on attribution problems, theoretical assumptions, sampling methods, and vocabulary studies.

229. Pressman, A. Common usage dictionary (English-Russian and Russian-English). The living language course, New York: Crown Publishing Company, 1958.

This course follows the method of Ralph Weiman. It contains 15,000 basic items and 1,000 essential items. Unfortunately, it does not state how the items were selected. It has glossaries of geographical and proper names.

230. Purin, L. A standard German vocabulary of 2932 words and 1500 idioms. Boston: D. C. Heath and Company, 1937.

This book contains 2932 alphabetically arranged words, 2000 derivatives, and 1500 idioms. It is for use in high schools, and in elementary and intermeidate courses in college German. It is based in part on: the Wadepuhl-Morgan (American Association of Teachers of German) Dictionary and the New York State Basic German Work List (1934) and the German Idiom List of C. D. Vial (SUNY

~~267~~ 268

230. (continued)

1933). The list includes 967 of the most frequently used words given in the Wadepuhl-Morgan Dictionary. The first 500 words in the Purin list include 400 found by Ortman as common to 12 of the word lists he examined. There are other words which were added to previous word lists based on recommendations of experts. They are the so-called "useful" words similar to those used in Fundamental French and other similar compilations of basic or first-stage (level) language vocabularies. English translations are provided for the German idioms. Semantic meaning is explained by examples of the most common meanings of each idiom. These examples are, in fact, sentence frames. English cognates are also provided where appropriate.

231. Rapoport, Anatol The stochastic and the 'teleological' rationales of certain distributions and the so-called principle of least effort. Behavioral Science, 1957, 2, 147-161.

Criticism of Zipf's principle and interpretation of the Simon and Mandelbrot derivations of the word frequency distribution function.

232. Reed, David W. A statistical approach to quantitative linguistic linguistic analysis. Word, 1949, 5, 235-247.

"The two elementary statistical devices presented are those which may aid in answering the following questions in quantitative linguistic analysis: (1) How much evidence should be collected in order to make a valid analysis of the frequency of linguistic forms? (2) When may quantitative differences in linguistic material be considered significant?" The two devices are the "Standard Error of Proportion" and the "Standard Error of Difference".

233. Richards, Jack C. A psycholinguistic measure of vocabulary selection. Paper presented at the annual meeting of the Canadian Linguistic Association, York University, Toronto, June, 1969, Eric Accession No. ED-035-860.

"Several basic problems in the field of the selection of vocabulary for teaching English as a foreign language are discussed.

233. (continued)

The nature of word frequency and word availability are considered, along with their limitations as measures of the usefulness of concrete nouns. Word familiarity is proposed as a psycholinguistic measure for noun selection, and some experimental evidence presented to demonstrate its validity. This is a preliminary report of a study which updates the 'general service list' of Michael West through establishing word familiarity figures for some 5000 nouns as well as updated frequency figures for written and spoken English."

234. Roberts, A. Hood A statistical linguistic analysis of American English. The Hague: 1965.

The author presents "a quantitative analysis of the segmental phonemes of a speaker of a North Central US idiolect. With the aid of a digital computer, the 10,000-word corpus was analyzed with results that should help fill several needs in present-day linguistic study. Among these findings are the following: (1) the etymological composition of English according to proximate sources by thousands of frequency; (2) the canonical forms of the words in the language according to the classification of the phonemes as vowel, consonant, semivowel and as to place and manner of articulation; (3) the frequency of occurrence of the phonemes of the language; (4) the average word length in phonemes and in syllables by thousands of frequency; (5) the relationship between the alphabetic and phonemic systems of notation; (6) the frequencies of occurrence of initial, intervocalic and final consonants and consonant clusters; (7) the entropy of English determined by the relative frequencies of the phonemes in the corpus and by word length in phonemes and in syllables; (8) the transitional probabilities of phonemes; (9) the Standard Error of a Proportion, the Standard Error of Difference between the two proportions, and the Standard Error Deviation for consonants and vowels separately and together."

235. Robinson, W. P. Cloze procedure as a technique for the investigation of social class differences in language usage. Language and Speech, 1965, 8, 42-55.

"Cloze procedure was used to investigate the nature and extent of the differences in verbal behavior of working and middle

235. (continued)

class boys. Words were deleted in sentences taken from 'formal' and 'informal', middle and working class letters and from middle and working class oral utterances. The results showed that the middle class boys used a wider range of words and preferred different words in this situation. The working class boys showed more conformity in their responses than the middle class boys, especially for the written materials. Fruitful lines for further research on 'restricted' and 'elaborated' codes are discussed."

236. Rodriguez-Bou, I. Recuento de vocabulario Espanol (Spanish vocabulary count). Rio Piedras, Puerto Rico: University of Puerto Rico Press, 1952.

The word count was encouraged and aided by the Organization of American States and UNESCO (program for Fundamental Education). It is a list of words in the Spanish language in accordance with the frequency of usage. It refers to Buchanan's Graded Spanish Word Book of 1929 which had 1,200,000 words. This word book considered more than 7,000,000 running words, covering both written and spoken Spanish of children and adults. It consists of Volume 1 and 2, Parts 1 and 2. Part 1's sources were newspapers and magazines, radio programs, religious works, and scholarly texts. It gives frequencies of lexical units of each source separately, based on 1,000,000 units. It includes frequencies of inflectional forms as well as of the head word. The introduction gives a fine history of word counts and their importance. The first list is of the 10,000 lexical units most frequently used in order of rank with separations at each 500 through the first 5,000. The second list gives the 20,000 most frequent inflectional forms in order of frequency rank. The third list gives the first 10,000 lexical units (same as list 1), in alphabetical order but with the frequency and frequency rank indicated. The fourth list gives the 20,000 word inflectional list (list 2) in alphabetical order with frequency and frequency of less than 16, listed in alphabetical order. Appendix A provides the methods, techniques and procedures used in compiling the frequency counts. Words counted include separately all the variations in the form of each word, including idioms. However, different semantic meanings of a word were not included, which keeps an otherwise excellent word count in an incomplete form. All words were included except unintelligible ones, invented ones, words without meaning, and some peculiarities of speaking or writing of children. Neologisms and regionalisms were included but not marked as such, if they were accepted by a panel of experts. Other words in current spoken or written

236. (continued)

Spanish, but not in dictionaries, were accepted if they appeared to be derived according to the laws of composition and derivation of words and were used by a good number of educated people. It was found that the first 105 words accounted for 50 percent of all words used. Data sources which were used were: oral vocabulary, associations, written compositions, and the count of Rodriquez and Casanova (University of Puerto Rico).
Oral: School children were placed in various situations at school and asked to write on the subject or discuss it orally in front of a tape recorder. From grades 1 to 6, some 1,073,245 running words were compiled.
Association (Controlled): Groups of nine words were used up to a total of 10 groups (90 words). Children wrote words which occurred to them from words of the nine in a group.
Association (Free): Students wrote all the words occurring to them in five minutes from association. Some 926,404 running words were obtained.
Written Composition: The procedure followed was that of Rinsland. The sample was grades 2 to 6 of public schools (803,622 running words). To these were added 586,141 compiled by Rodriquez and Casanova in a similar collection exercise.
Recognition Vocabulary: Newspapers from January through June, 1947; some 91 editions on alternate dates. (1,050,000 running words).
Radio programs of various types (465,600 running words).
Religious (all types in Puerto Rico).
Buchanan's Graded Spanish Word Book (1,300,000 words).
Basic Educational Texts for Elementary Schools.
Books of supplemental reading for elementary schools.
Appendix B - The first, most used, 105 words.
Appendix C - Book references.
Appendix D - Procedures for obtaining compositions for children.
Appendix E - Free association stimulation words.
Appendix F - Buchanan's 40 categories of written materials.
Appendix G - Directions for arranging order in lists of lexical units and inflectional lists.
Volume 2, Part 1 starts with a summary of the introduction and history of word counts. It gives the 20,542 lexical units of the count with their inflectional forms, as line items. Under 14 columnar headings are listed the frequencies of appearance of the lexical units by source of the words (A-H). Part 2 continues Part 1 for words beginning with I-Z. It also contains Appendices A-G as in Volume 1.

237. Rose-Innes, A. Fundamental spoken Japanese. (Revised and enlarged by W. Kos, S. J. and also known as A new 3rd edition of conversational Japanese for beginners). Tokyo: Meiseisha Publishing Company, 1967.

> Part 1 is a graduated exercise in conversational Japanese developed by Japanese in Japanese. An English translation is given for each. Part 2 is elementary grammar of spoken Japanese. Emphasis is placed on the sentence, not on the words. Part 3 is an explanatory vocabulary of common Japanese words. This is an update of an earlier publication by the same author which printed the vocabulary separately.

238. Russo, G. A., A combined Italian word list, Modern Language Journal, 1947, 31(4), 218-240.

> "This list contains 3,173 Italian words with indications of their relative difficulty. The collection was compiled from two earlier selections: (1) the 'Knease List' of some 400,000 running words based on 40 Italian literary works published in Italy and scored according to range and frequency, and (2) the 'Skinner List' constructed according to range only from the Italian-English vocabularies of 45 Italian textbooks published in the United States."

239. Rutherford, R. W. and M. Wears, ed., Enquete sur le language de l'enfant Francais, (Investigation of the language of French Children): The spoken language of nine-year-old French Children. The Nuffield Foundation (Leeds, England), 1969.

> "Transcriptions of recorded conversations of nine-year-old French children are analyzed and presented in this comparative word count. The actual count of the 55,588 word corpus is arranged alphabetically and contrasted with selected, identical words found in the Francais Fundemental word list. Proper nouns are listed separately at the end of the regular count, and grammatical functions of items are mentioned when a word appears. The count lists word frequency, word totals, and the Francais Fundemental word count. Discussion of classification, column labels, and notes on child language are included."

240. Rosman, E. *A pilot project on vocabulary selection for foreign students.* New York: Teachers' College, Columbia University, November 1962.

This work describes a method of making word frequency counts for Afghan students at a US university. A set of 30,000 cards of words and phrases is available at the International Center of Teaching Materials at Teachers' College. It emphasizes reading rather than oral vocabulary. It includes a 3,436 word count and a list of 103 of them with the greatest frequency.

241. Sacleux, C. *Dictionaire Français-Swahili (French-Swahili dictionary).* (2nd edition revised and augmented) Paris: Institute of Ethnology, 1959.

Sacleux was a missionary who worked principally in Zanzibar. Unfortunately, the dictionary does not give any indication of the techniques of word selection or other procedures used in its preparation. The previous edition (1939) had 1115 pages (but printed in different font). There is, however, an introduction which indicates the geographical distribution of Swahili, the alphabet used in the dictionary, the Swahili dialects, and notes that the Zanzibar dialect (Ki-Ungudya) was the principal basis for the dictionary and had preference in its development. The dictionary also contains a two-page bibliography.

242. Savard, J. G. *Analytical bibliography of language tests (Bibliographic analytique de tests de langue).* Quebec: Les Presses de l'Universite Laval, 1969 (bilingual).

This recent book is divided into an introduction and seven parts. The first five parts are subdivided into an index of titles, index of authors, and analytical. The seven parts are: Second Language (150 tests), Mother Tongue (National Language-150 tests), Bilingual Tests (English-Spanish and English-French), Aptitude Tests, Psychological Tests, Miscellaneous (not in the first five parts), and Index of Publishers. Parts 1, 4, and 5 are applicable to student selection and training.

243. Savard, J. G. *La valence lexical (word coverage).* Paris: Didier, 1970 (in French).

The purpose of the book is to develop an alternative to frequency of occurrence as an objective measure by which to select words

243. (continued)

for teaching vocabularies, especially for beginners in a second language. Once developed, it could also be used to refine existing vocabularies. Source documents were Fundamental French, 1st and 2nd Levels, dictionaries of synonyms, and analogical dictionaries. The proposed alternative to frequency in establishing vocabularies is valence. The criteria for valence are stated as *definition* (explaining power, or how often the words can be used to explain or define others; in other words the basic nature of the word), *inclusion* (the number of words for which it is a synonym and for which it may be substituted), *combination* (combining power. Its use in compound words and idioms), and *extension* (semantic range or power; the quality of having more than one partially or completely different meaning. In the analysis each criterion was given equal weight and its computed value was added to the other three to arrive at the total valence of the word in question. The text describes the proceudres used and is plentifully supplied with appendices (4), tables (16), and figures (6). Of interest is the fact that the basic French vocabulary so derived has 897 words which in number approximates the 850 of basic English. However, when valence is computed, there appears to be no correlation between words rating high on valence and those on frequency so more research is required to determine the best basis for vocabulary construction.

244. Scholes, Robert J., On functors and contentives in children's imitations of word strings. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 167-169.

"Young children (mean age 3 years, 11 months) were asked to repeat word strings presented from tape. The strings varied in length from three to five words; in sentencehood in that some were well-formed sentences, some were anomalous, and some were syntactically diviant; and in word types in that some strings contained all real words, some contained real function words plus nonsense items, and some contained real content words plus nonsense items. The results of this experimentation suggest that children's differential imitation of contentives and functors is accounted for by an 'identify and retain contentives' strategy and that the principal criteria for the classification 'contentive' are phonological form and semantic function."

245 . Schonell, Fred J., Meddleton, Ivor G., and Shaw, B. A. *A study of the oral vocabulary of adults: an investigation into the spoken vocabulary*

245. (continued)

of the Australian worker. Brisbane: University of Queensland Press, 1956, and London: University of London Press, 1957.

> The work contains a brief account of the history of word frequency studies (pages 18-27) along with a more particular discussion related to investigating vocabulary size. Spoken samples of the speech of Australian workers were taken by interviews, recordings on the job and in public places. The author discusses utility yields of methods and correspondences of the hand tabulation of a total of tokens accounted for by 85 words, all of which were words "..used to promote the flow of speech rather than to inject meaning into what is said."

246. Seashore, Robert H., and Eckerson, Lois D. The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, 1940, *21*, 14-38.

> This study presents a detailed account of work aimed at discovering the size of vocabularies by means of multiple-choice tests derived from dictionary entries.

247. Sebeok, Thomas A. (ed.) *Style in language*. Cambridge, Mass., New York, and London: MIT Press, 1960.

> A collection of papers presented at a conference on stylistics held at Indiana University in 1958, by literary critics, linguists, psychologists, and cultural anthropologists working in stylistics.

248. Shannon, C. E. Prediction and entropy of printed English. *Bell System Technical Journal*, 1951, *30*, 50-64.

> "A new method of estimating the entropy and redundancy of a language is described. The method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of the ideal predictor are developed."

249. Shapiro, B. J. The subjective scaling of relative word frequency. Doctoral Thesis, Harvard Graduate School of Education, 1967, Ann Arbor, Michigan: University Microfilms, 1972.

This study explores subjective scaling of word frequencies and its relationship to objective scaling. If a relationship could be established, the enormous samples required in objective scaling studies could be avoided by using subjective scaling and converting it mathematically to the equivalent objective scaling result. The study explores the subjective scaling of relative word frequency of English in relation to the Fechnerian (logarithmic) and Stevens (power) psychophysical theories, various informant populations, and both the written and the spoken language. The Thorndike-Lorge 30,000 word Teacher's Word Book (1944) and the Francis and Kucera word count of 1965 were used as the objective criterion measurements. Eighty-eight word-stimulti for the subjective scaling were selected from these sources to cover frequencies ranging from .2 to 68,000 per million. 184 informants were selected from as varying populations as sixth and ninth graders, college sophomores, and adults in as widely distributed occupations as industrial chemistry, elementary school teaching, and newspaper reporting. Two subjective scaling methods were used: multiple rank order and magnitude estimation. In addition, half of each informant group responded in terms of written language and half in terms of spoken language. The author concludes that the magnitude estimation technique tended to follow the Stevens (power law) model and the multiple rank order technique was closely related to the Fechnerian (logarithmic law) model. Chi Square tests showed, however, that the observed multiple rank order data did not fulfill all the assumptions of its analytic technique. In addition, the magnitude estimation and multiple rank order techniques were logarithmically rather than linearly related. Shapiro also found that relative word frequencies are a prothetic (how much?) psychological-additive variable rather than a metathetic (what kind or where?) substitutive variable. He further states that the same observations apply to other linguistic units, such as syllables, grammatical constructions, and letters. He concludes that they are best measured subjectively by the magnitude estimation technique, but that there is a need for additional studies to verify his findings. The study is profusely illustrated by tables and figures and contains samples of forms used in sample collection from informants. In addition, the procedures used are amply detailed. This is an important contribution to ongoing work on alternative methods of obtaining word frequencies without the voluminous sampling used in the objective or direct counts.

250. Sherwood, John, and Horton, Iver Phoneme frequencies in Australian English: a regional study. Journal of the Australasian Universities Language and Literature Association, 1966, 26, 272-302.

> This study involved transcribing 16,800 running words from the speech of twenty-two adults and thirteen children according to the phonetic scheme developed by Professor A. G. Mitchell for Australian English. The statistical results are compared with other published counts. The work is part of a larger study of regional variation in Australian English.

251. Siliakus, H., and Morris, K. Some reflections on the lack of accuracy of word frequency lists. Review of the ITL, (Institute of Applied Linguistics, Louvain, Belgium), 1970, 9, 11-18.

> The article discusses the ubiquitous problem of error in frequency counts due to low occurrence of important items in word counts taken from relatively small lengths of running text, and illustrates with examples. They reinforce the argument for longer samples of running text and shorter frequency lists, i.e., thematic or topical.

252. Simon, Herbert A. On a class of skew distribution functions. Biometrika, 1955, 42, 425-440.

> This article considers a variety of such functions; in discussing empirical distributions, he examines the distribution of word frequencies.

253. Skinner, B. F. The distribution of associated words. Psychological Record, 1937, 1, 71-76.

> Skinner shows that the rank-frequency relation described by Zipf applies to "samples of speech selected on a semantic basis". He illustrates his thesis by an examination of free association responses.

254. Society for International Cultural Relations (Japan) Japanese basic vocabulary. Tokyo: Kokusai Bunka Skinkokai (KBS), 1941.

> This is a 2000 basic word list. The class and inflections of each word are given, and the several meanings of each word are carefully

254. (continued)

explained. With its examples of compounds and synonyms, this is almost a basic dictionary. The words were selected by a committee based on subjective criteria. One of the purposes of this book was to assist in teaching Japanese to a foreigner.

255. Society for International Cultural Relations (Japan) KBS bibliography of standard references for Japanese studies with descriptive notes. Tokyo: Kokusai Bunka Shinkokao (KBS), 1961, 4.

This bibliography is similar to, but extends further back into history than the one by Yamigawa. It has 16 chapters of which the following are especially important bibliography, dictionaries, phonetics and phonology, grammar, and special languages and lexicology.

256. Somers, H. H. Analyse mathematique du language-lois generales et mesures statistiques. Louvain: 1959.

In this book after examining the formulas of Zipf, Mandelbrot, and Simon, the author proposes a lognormal distribution for vocabulary and applies it to texts for which counts have been published. He uses it in estimating vocabulary, text-length ratios, word-length distributions, etc. A formula for type-token rations is given.

257. Spolsky, Bernard et al, A spoken word count of six-year-old Navajo Children. New Mexico University, 1971.

"As part of a study of the feasibility and effect of teaching Navajo children to read their own language first, a word count collected by 22 Navajo adults interviewing over 200 Navajo 6-year-olds was undertaken. This report discusses the word count and the interview texts in terms of (1) number of sentences, (2) number of words, (3) number of tokens, (4) type-token ratios, and (5) word-length. A frequency list gives all words used by at least 2 children. The words, mostly Navajo, are grouped in order of frequency and in alphabetical order with each frequency. A supplement lists, alphabetically, all words from the interview texts (whether used by children or adults). Frequency and range data for adults and children are given separately and in total for each word.

258. Stone, P., *et al*. *User's manual for the general inquirer*. Cambridge, Mass.: MIT Press, 1968.

> This is a companion volume to *"The General Inquirer: A Computer Approach to Content Analysis"* (MIT Press, 1966). The purpose of this book is to provide technical specification of the computer programs in the General Inquirer content analysis system and detailed instructions for using these programs. The manual is divided into four parts of 3-6 chapters each and a series of appendices. Part 1 is an introduction to the system and specifications for preparing dictionary and text data. Part 2 discusses primary programs for assigning tags to text, and performing operations of listing, counting, and retreival on tagged text. Part 3 relates secondary programs for processing tag scores. Part 4 gives two types of secondary programs to facilitate the development of dictionary categories--one for generating a key-word-in-context index of a sample of text, and the other for displaying a dictionary in a special "cross-sorted" format.

259. Swenson, E., and West, M. *On the counting of new words in textbooks for teaching of foreign languages*. (Bulletin No. 1, Department of Educational Research) Toronto, Canada: University of Toronto Press, 1934.

> This short study is an excellent background on the subject of word counting. It is in two parts: on counting of new words and the history and purpose of word counting and analyses its procedures (as of 1933-1934). Of special interest are the chapters on the origin and counting of a speaking vocabulary. Part 2 on rating scales provides specific methods for rating the difficulty of learning meaning, idioms, cognates, compounds, and spelling-pronounciation discrepancies (i.e., the words that do not sound the way they look or vice versa). The study was undertaken to eliminate some of the confusion of the techniques of word counting and vocabulary control. It discusses Palmer's work in the field, reasons for word counts, differences between written and spoken counts, methods of rating words, word rating scales, and testing the scales for reliability and intelligibility.

260. Swenson, Rodney A frequency count of contemporary German vocabulary based on three current leading newspapers. *Dissertation Abstracts*,

260. (continued)

1967, *18*, 222A-23A (Also final report of the Director on a frequency count of contemporary German vocabulary based on three current leading newspapers. A project of the US Office of Education in cooperation with Hamline University, St. Paul, and the University of Minnesota, Minneapolis. Washington, D.C., Department of HEW, USOE, 1967.)

The foreword discusses the lack of validity of earlier word counts because of changes in language and changes in the goals and objectives of teaching foreign languages. The goal now is oral, which leads to the oral vocabulary needed to learn the most frequently used words in current speech. The newspapers selected for the study were: *Die Welt* in Hamburg (183,840 words), *Suddeutsch Zeitung* in Munich (232,280 words), and *Frankfurter Allgemeine Zeitung* in Frankfurt a/M (167,700 words). The papers were selected since they all had a fairly well-educated reading audience with wide West German geographic spread. Each had a circulation of 210,000 to 250,000. Since in Germany two-thirds of the population read newspapers, newspapers were selected for the sampling as they were considered to contain words in general usage. As noted above, the sampling did not include East Germany, Austria, or Switzerland. The sampling took place from 1 October 1964 through 31 January 1965 (four months). 584,000 running words of an estimated 12 million were tabulated. Samples were taken from columns at least six inches in length. Every fifth column was selected and the first 120 words counted and tabulated. Advertisements, want ads, headlines, and picture captions and supplements were not used. Proper nouns, geographic names and abbreviations were also excluded from the count but were tabulated and filed separately. Word forms were tabulated under the root form or infinitive. Tabulations followed Pfeffer's Basic Spoken German. The final lists are of the first 500, 1000, and 6500 words of the German language as reflected in newspapers. The tabulations indicate the count by newspaper and by total. The conclusion is that with the 1500 words in the longest list, one could understand a considerable amount of contemporary German. The list does not include grammatical forms and is religious, sports, political, literary, and science oriented. Frequencies indicate changes are required in the sequence of instruction, especially in verbs, since forms used now are not being taught first. Articles (grammar) have about the same frequency now as in the counts of the 1920's. High frequency words in one paper were generally high in the others. This was not so true of low frequency words (probably reflecting newspaper policy or regional difference). There are two major lists: numerical frequency (one for each newspaper and a total) and alphabetical (for the total count only).

261. Tadashi, Kikuoka The 1000 most important Japanese newspaper character compounds in order of descending frequency. South Orange, N. J.: Institute for Far Eastern Affairs, Seton Hall University, 1965.

> The sources are: A glossary of journalistic terms. Tokyo: Nihon Simbun Kyokai, 1961, and the Dictionary of journalistic terms. Tokyo: Asakai Shimbun-sha, 1961. The original selection of compounds was made by Dr. Hirosha Okube at Hosei University, Tokyo. It was updated by the present author.

262. Tannenbaum, Percy H., and Williams, F. Prompted word replacement in active and passive sentences. Language and Speech, 1968, 11, 220-229.

> "A conceptual focus formulation developed in a prior study of encoding of active and passive sentences led to predictions concerning how such sentences are stored and their main semantic units are retrieved from memory. The formulation posed a dominant subject-verb linkage in active sentences but an object-verb linkage characterizing passive sentences. Individuals were presented with either the subject, or verb, or object of previous exposed sentences and were required to replace the other two missing words. As anticipated, the subject was a better prompt for the verb in active than in passive sentences but the reverse relationship obtained when the object was the cue. Similar predictions for situations when the verb was the prompt word were supported when the subject was the response but not when the objects were to be replaced."

263. Tarnoczi, Lorant. Wortbestand, wortschatz, wortfrequenz. IRAL, 1971, 9, 297-318.

> Written in German, this article criticizes the manner in which minimal lexical inventories are made and questions the claims made for such lists. He refutes the claim that a minimal vocabulary of 1000 to 2000 words will enable one to understand 75 percent of a text in a given language, particularly since the texts used are theme oriented and this fact is not taken sufficiently into consideration. The author proposes that minimal vocabularies based on frequency counts must distinguish between basic vocabulary and thematic vocabulary and that a minimal vocabulary must vary according to the teaching objectives.

264. Taylor, G. Learning American English. New York: Saxon Press, 1954.

This book is planned to meet the needs of adult students at the beginning of intermediate states of learning English as a second spoken language. The English described is a basic, informal, spoken language, used by the majority of US citizens. There are 17 lessons each followed by exercises. There are over 1500 English words used, but the author suggests concentrating on the first 550. Word lists are derived from Thorndike and Lorge's 30,000 Word Teachers' Word Book (1944) and the KLM List of Bongers' History and Principles of Vocabulary Control (1947).

265. Thomson, Godfrey H., and Thompson, J. Ridley Outlines of a method for the quantitative analysis of writing vocabularies. British Journal of Psychology, 1915, 8, 52-69.

The question addressed by the study is: "How can we find a measure to enable us to estimate the total vocabulary from the study of a sample?" (page 58). The method they develop consists of assigning weights to particular words according to the number of times they appear in the sample; the proposed formula is then used to estimate the author's total vocabulary. The technique is tested through a study of chapter fifty-five of David Copperfield.

266. Thorndike, Edward L. On the number of words of any given frequency of use. The Psychological Record, 1937, 1, 399-406.

This is a discussion and criticism of Zipf's rank-frequency hypothesis, the essay includes data from a 4.5 million word sample of the language of children's books.

267. Thorndike, Edward L. and Lorge, I. The teachers' word book of 30,000 words. New York: Teachers' College, Columbia University, 1959.

This publication consists of five parts and is the last of a series of word lists published by Thorndike with or without co-authors, e.g., the 20,000 word lists. Part 1 is a list of words occurring at least once per 1 million words. Part 2 is a list of words occurring at least once per 4 million words. Part 3 is an explanation of Parts 1, 2, 4, and 5. Part 4 includes the number of occurrences of words occuring 1000 or more times in either of the counts (Lorge Magazine Count and Lorge-Thorndike Semantic Count). Part 5 is a list of 500 words occurring most frequently and of the 500 occurring next most frequently.

268. Uhlirova, L. On statistical experimenting in syntax, Statistical Methods of Linguistics, 1969, 5, 18-33.

"The first stage of statistical research on Czech word order in relation to syntax and to the so-called topic-comment bipartition of the sentence. A necessary prerequisite of this research is the determination of whether or not the linguistic material in question is representative from the statistical point of view. Hence, the objective is to establish by statistical experimentation whether or not a certain corpus is representative (sufficient) for the purpose of a synatactical word-order analysis. Samples of 1000 successive clauses were selected arbitrarily from texts of different genres and analyzed syntactically in accordance with Czech authorative grammar. Three classes of clauses were differentiated: clause simple sentence (J), sentence (V). Tables are included which indicate the cumulative distribution of clause length across clause types and by clause type, frequency of occurrence of each clause type, frequencies of syntactic categories in clauses of varying lengths, as well as graphs representing these data."

269. University of Michigan (English Language Institute) Selected Articles from language learning. (Series 1. English as a foreigh language) Ann Arbor, Mich.: Langugae Learning Reprints, 1953.

This is a selected group of articles by prominent scholars of English. The articles were picked to emphasize the "new" (1953) approach in structural linguistics from items on word and sound and from the system of contrastive patterns in which the items operate. There are six parts to the collection, each with from 3 to 13 articles, They include language learning, language teaching, grammar, pronounciation, vocabulary, and testing.

270. University of Michigan (English Language Institute) Teaching and learning English as a foreign language. Ann Arbor, Mich.: University of Michigan Press, 1962.

This work is a basic explanation of English as a foreign language. It includes an introduction on adult learning, sounds, structure arrangement and form, words and vocabulary, and contextual orientation.

271. Vakar, Nicholas P. A word count of spoken Russian--the Soviet usage. Ohio State University Press, 1966. (This is the final report on US Office of Education HEW Contract CEC-3-6-062046, July 1969)

> There are two parts (volumes) to the report; Part 1, vocabulary (normal colloquial vocabulary)--obtaining samples of colloquial Soviet Russian speech is difficult. Professor Vakar notes that since 1956, however, Soviet drama has come to deal with everyday problems and has been presented in the language of the audience. As a result, Professor Vakar based his study on an actual count of 10,000 words from 50-word samples taken from 200 acts of 93 plays published since 1957. With the small sample, Professor Vakar assumed that the most common words occur in virtually every conversation of any length, so the sample need not be large. In fact, he found that 360 words of a total of 2360 words in the 10,000 word sample represented 73 percent of all occurrences and are satisfactory for intelligent adult oral communication. Also of note are some 75 word-clusters which indicate the cumulative frequency of occurrence of certain stem or roots. Part 2, sentence structure (colloquial in Soviet usage). There is a tremendous difference between literary and spoken Russian. A set of a few hundred common words, grammar fundamentals and favorite turns of phrase constitutes the core of ordinary conversation. Chapter 2 gives the basis for the sentence sample. It also used monologues and dialogues of 93 plays written from 1956-1964, representing a statistical universe of 1 million running words. From these 1000 sentences were randomly selected. Sentences tend to be short--1 to 5 words (75 percent). Also included is a glossary of nouns, verbs, adjectives, and adverbs as well as four appendices.

272. Vakar, Nicholar P. Statistical methods in the analysis of Russian. Slavic and East European Journal. 1967, 11, 59-65.

> Vakar here contrasts two counts, one based on a word count of contemporary plays and the other on a word count of actual conversations. He then discusses the implications of such counts for language teaching.

273. Van den Eynde, R. Grammaire Swahili (Swahili grammar). Brussels: Waithhoz-Legrande, 1944.

> This book is in French. It does contain a vocabulary as well as the grammar suggested by the title. This book was originally intended for students who would later spend some time in the (then)

273. (continued)

Belgian Congo. Although the Swahili dialect spoken in the Republic of Zaire is a poor one found largely in Katanga (Kingwana), the author has addressed himself to "pure" or standard Swahili of the east coast of Africa. The grammar part takes up the first 88 pages with nine chapters. It is followed by the French-Swahili vocabulary. The alphabetized French-Swahili vocabulary is followed by a special vocabulary on units of measurement, days of the week, and a Swahili-French vocabulary.

274. Vander Beke, George E. *French Word Book*. (American and Canadian Committees on Modern Languages) New York: The MacMillian Company, 1929, 15.

The basic method that was used was that of Henmon in his *French Word Book Based on a Count of 400,000 Running Words* (see above) where categories of words were selected and teachers of French were used to help compile the ultimate product. The Committee combined their list with that of Henmon. The results was a count of over one million running words from eighty-eight examples of French prose. The study, however, did not include anything other than printed material and used no particular criteria for selection of sources. The texts used to draw up this count were selected from nineteenth and twentieth century literature, and were divided into twelve categories. The method of tabulation and analysis is also described. Part I lists the words omitted from the count but listed in the Henmon study. Part II lists the words by range and gives both the range and frequency. Part III combines the list with the Henmon list and gives the range, frequency, the Henmon frequency and the total frequency of the word. The order is alphabetic. An Appendix list those words which Henmon listed, but which the committee found to be too low in frequency to be counted.

275. Van Spaandonck, M. *Practical and systematic Swahili bibliography-linguistics 1850-1963*. Leiden, Netherlands: E. J. Brill, 1965.

Chapter headings indicate the coverage by classification: general, linguistics (including grammars, instruction, and phrase books, supplemental linguistic studies, dictionaries and vocabularies, supplemental vocabulary studies), literature, Katanga Swahili, and an appendix.

276. Voelker, Charles H. The one-thousand most frequent spoken words. *Quarterly Journal of Speech*, 1942, *28*, 189-198.

276. (continued)

Voelker's list is based on a sample of 99,400 words gathered from the speech of older adolescents.

277. Walsh, S. (ed.) English language dictionaries in print--a comparative analysis. Newark, Delaware: Reference Books Research Publications, Inc., 1965.

This compilation compares dictionaries by cost, number of entries, age, group suitability, date of publication, and user ratings. It also describes each one briefly.

278. Wepman, J., and Hass, W. A spoken word count (children 5-7). Chicago: Language Research Associates, 1969. (See also A spoken work count by Wepman and Jones)

This study attempts to supply additional information on the quantitative aspects of childrens' word usage. It also discusses uses of childrens' word counts aside from inclusion in readers. The sample group was 90 middle-class English-speaking metropolitan children aged 5-7 (30 each) and boys/girls (45 each). The procedure was similar to that used by Wepman and Jones for adults. The material was computer processed into three lists: word frequency, in order of frequency, part of speech by grammatical class, and alphabetical list of all words used by at least two speakers. It also contains a short bibliography.

279. Werner, H., and Kaplan, B. An organismic-developmental approach to language and the expression of thought. Symbol formation, New York: John Wiley and Sons, Inc., 1963.

The authors use their perspective of psychological phenomena to demonstrate how that perspective enables an individual to order and integrate data on symbolization and language behavior. The book contains five parts: organismic-developmental approach, formation and general changes in verbal symbolic behavior in the course of ontogenesis, processes which underlie the primordial states of linguistic representation through study of adult behavior, linguistic representation under differing conditions of communication, and symbol formation in non-verbal media.

280. West, M. On learning to speak a foreign language. London: Longmans, Green, and Company, Ltd., 1933.

> This is a book on teaching spoken English as a foreign language. It discusses purpose (aim), policy (theory), techniques (methods), vocabulary (reading and speaking vocabularies), and minimum adequate vocabulary to include the concept of completeness and vocabulary design. West considers the Thorndike and Horn 1000 most common word lists, the American College list (1000 words, and similar to Thorndike's) the 1000 words used by the Adult Education Society of New York, Palmer's list of 600 words, Palmer's Composite Word Frequency List of 1000 words, and Ogden's Basic English Vocabulary. He finally arrives at a list of 996 words and procedures for structuring lessons from those words and his text.

281. West, M. Definition vocabulary. (Bulletin No. 4, Department of Educational Research, Ontario College of Education, University of Toronto) Toronto, Canada: University of Toronto Press, 1935.

> This is a study on how to determine the vocabulary to be included in a dictionary for foreigners. The author argues that the major problem in preparing a dictionary for foreigners learning English is selecting the words to define the words in the dictionary, in such a way that both will be understood. West solves the problem by determining some 1490 words and 85 irregular verb forms and plurals with which he can explain all the words and idioms in the proposed dictionary of 17,727 words and 6,171 idioms. (This, of course, becomes difficult when the idioms are not self-explanatory on the basis of individual word meaning.) The book has three chapters and 28 explanatory tables.

282. West, M. A general service list of English words. London: Longmans, Green and Company, Ltd., 1960.

> This book contains a frequency count based on 5 million words and has a semantic count by percentages for written and printed English.

283. West M. Teaching English in difficult circumstances (teaching English as a foreign language). London: Longmans, Green and Company, Ltd., 1960.

283. (continued)

This book is based on experience in teaching English in India and the Middle East. In general, the book is a teachers' guide but it has an appendix with a minimum adequate (1200 word) vocabulary for spoken english and a classification guide to accompany it.

284. West, M. *An international readers' dictionary*. London: Longmans, Green and Comapny, 1965.

This dictionary contains 24,000 items (18,000 words and 6,000 idioms). It is designed for the use of persons for whom English is other than their native language. It supercedes and updates the New Method Dictionary by West and Endicott (1935-1960). Explanations are made within a vocabulary of 1490 words held to be among the most common in the English language as learned by foreigners. It excludes scientific and technical terms in common use in news media and books. It excludes also certain derivations and compounds when their meanings can be inferred from the root word and context.

285. West, M., and Bond, O. *A grouped-frequency French word list*. Chicago: University of Chicago, 1939.

The purpose of this book was to re-work the Vander Beke *French Word Book* into forms more useful to teachers. This book by West and Bond has three parts: frequency list in numerical order with inflectional forms under head-words and listed in 100 word groups, index--an alphabetical list of head-words, and two appendices: fifty Latin roots and common French affixes (prefixes and suffixes).

286. Whatmough, J. *Language--a modern synthesis*. New York: Mentor Books, 1956.

In many respects this book looks like a companion piece to Zipf's *A Psycho-Biology of Language.* It has 13 chapters including past and present languages, words and meanings, the uses, structure, analysis, and neural basis of language, as well as the mathematics and statistics of language.

287. Whiteley, W. *Some problems of transivity in Swahili*. London: Luzac and Company, Ltd. (for the University of London), 1968.

287. (continued)

Two areas in which existing Swahili dictionaries are weak are transivity and verbal extensions. This book deals with transivity (of verbs). Transivity deals with the various relationships which obtain between a verb and a noun or nouns to which the label "object" is often accorded. This book deals with the subject in detail, as an aid to writing more precise meaning into dictionaries.

288. Whiteley, W. Swahili--rise of a national language. London, England: Methuen and Company, Ltd., 1969.

This book is a broad survey of the Swahili language and literature, its early history, spread, status in the colonial period, its current status, and its prospects. One chapter is devoted to "standard Swahili" and a bibliography is included.

289. Whiteley, W., and Gutkind, A. A linguistic bibliography of East Africa. Kampala: East African Swahili Committee and East African Institute of Social Research, Makere College, 1958.

This is a classified bibliography indexed in part by country (Tanzania, Kenya, and Uganda), and in part by language (Swahili). It is very useful in finding local names for flora, fauna, as well as for more general works on the languages and linguistics of East Africa.

290. Wilson, P. English-Swahili (classified vocabulary). Nairobi, Kenya: East African Literature Bureau (undated).

The vocabulary is classified by vocation as: agriculture engineering, fishing, household, medical, and veterinary.

291. Wilson, P. Simplified Swahili. Nairobi, Tanzania: East African Literature Bureau, 1970.

This book is written for the individual who wants to achieve a quick general knowledge of Swahili. Grammar is kept moderately simple and is introduced as required and in order of relative importance of subject matter. It is an up-to-date book which will help individuals learn spoken Swahili. It includes translation exercises and keys to them. At the end of the book there are Swahili-English and English-Swahili vocabularies.

292. Winter, Ralph Dana English function words and content words: a quantitative investigation. Dissertation Abstracts, 1954, 14, 1084-1085.

> The author chose two texts of 4000 words each and parsed them according to the system established by C. C. Fries. Among the statistical measures employed were: word length in segmental phonemes, gaps between repetitions of the same word, "average interval between successive occurrences of a word", and "sum of the squares of the intervals between successive occurrences of a word". The study shows that quantitative data generally supports the division of function and content words.

293. Wisbey, R. A. (ed.) The computer in literary and linguistic research. (papers from a Cambridge symposium) Cambridge, England: Cambridge University Press, 1971.

> This is a compilation of articles on the subject in seven parts, each of which has from 3 to 6 articles. Parts are titled as follows: Lexicography, Textual Archives, and Concordance Making; Textual Editing and Attribution Studies; Vocabulary Studies and Language Learning (The most pertinent item is D. G. Burnett-Hall and P. Strupple's "The Use of Word Frequency in Language Course Writing"); Stylistic Analysis and Poetry Generation (The most important item is T. R. Tallentire's "Mathematical Modeling in Stylistics: its Extent and General Limitations"); Computer Applications to Oriental Studies; Problems of Input and Output; and Programming the Computer for Literary and Linguistic Research.

294. Wright, C. W. An English word count. (Department of Education; Arts and Sciences Research Series No. 15, National Bureau of Educational and Social Research, Praetoria, South Africa, 1965) London: Longmans, Green and Compnay, Ltd., 1965.

> This count is based on written English in South Africa. The counts were taken from The Bible, newspapers, periodicals, literary works, and correspondence. It has three lists covering 20,000 words: (1) first 1000 words in alphabetical order with an indication of the grouping of 100 words in which the word falls, (2) first 10,000 words, and (3) second 10,000 words.

295. Yamagiwa, J. (ed.) *Japanese language studies in the Showa period*. (Univeristy of Michigan Center for Japanese Studies, Bibliographic Series No. 9) Ann Arbor, Mich.: University of Michigan Press, 1961.

> This is a bibliography of modern Japanese work on the Japanese language since 1926 (The Showa Period). Chapter headings are: Bibliographies, Essay Series and Journal, Dictionaries, Encyclopedias, and Indices of Vocabulary, Outlines aand Description of Japanese Language Studies, History of Japanese Language Studies, Phonology, Grammar, Relationships of Japanese to other Languages of East Asia, History of the Japanese Language, Dialect Studies, Writing Systems, List of Publications, and Authors and Editors.

296. Yamagiwa, Joseph K. Linguistic data: some quantifications. *Studies in languages and linguistics in honor of Charles C. Fries*, Albert H. Marckwardt (ed.), Ann Arbor: 1964, 35-53.

> This study provides a statistical examination of the stylistic varieties in contemporary Japanese.

297. Young, I., and Nakajima, K. *Learn Japanese--college text*. (Asian Language Series) Honolulu: East-West Center Press, 1967, *1-4*.

> This series was originally written as *Learn Japanese-Pattern Approach*. The dialect used is of native speakers of a middle class background, college education, residents of the Yamanoto area of Tokyo and 25-45 years of age. The pattern approach is more than formula-application. It develops a new presentation based on association and repetition. It reflects connecting links between modes of utterances or patterns. A pattern is a structure related to other structures. Moving from one structure to another is done by transformation. The patterns reflect "live" situations as well as the structure of the language. The material is based on a contrastive study of English and Japanese structure. Volume 1 has 15 lessons. The general format of lessons includes: useful expressions, pattern sentences, dialogues, notes, vocabulary, Hiragana practice, and drills. In many respects, this approach parallels that of Jorden and Chaplin's *Beginning Japanese*, but teaches hearing, reading and writing, as well. Volume 2 is a continuation of Volume 1 but is more advanced. Volume 3 is an introduction to Kanjii characters. It has 15 lessons and seven appendices, including a glossary. Volume 4 introduces more Kanjii characters and has 15 lessons and seven appendices, including a glossary.

298. Zale, E. M. (ed.) (Proceedings of the) Conference on Language and Language Behavior. New York: Appleton-Century-Crofts, 1968.

The Conference on which this volume reports was held under the sponsorship of the Center for Research on Language and Language Behavior at the University of Michigan in 1966. Major topics discussed were: first language acquisition in natural setting, controlled acquisition of first language skills, second language learning, linguistic structure above sentence level, phonology and phonetics, and language impairment. Most of the six subject areas were covered by four speakers each. The three major addresses, attended by all conferences were as follows: (1) "Scylla and Charybdis, or the Perilous Straits of Applied Research: by A. P. Van Teslaar, (2) "Thought and Language" by James J. Jenkins, Director of Research, Center for Research in Human Learning, University of Minnesota, (not included in the report, but published otherwise by the University of Pittsburgh), (3) "Word Frequency Studies and the Lognormal Distribution" by John C. Carroll. Carroll's address is an extended edition of the one actually presented to the Conference. The main theme has also been published separately in several forms. Also of interest are the remarks on high and low association passages in the talk by Sheldon Rosenberg on "Language Habits and Recall of Connected Discourse", the discussion of negation in Japanese in "What Does a Child Mean When He Says No?" by David McNeill and Nobuke B. McNeill, "The Indices of Coverage: A New Dimension in lexicometrics by W. F. McKey and J. G. savard, "Auditory Discrimination and the Learning of Languages" (in French) by Emmanuel Companys, and "Remarks on the Predictive Value of Differential Analysis in Phonology" (in French) by Guy C. Capelle.

299. Zimmermann, Jon E. Word frequency in the modern German shorter narrative. Dissertation Abstracts, 1968, 28, 3362A.

The word count was based on 702 samples of narrative prose by 266 different authors; 160,000 words were randomly selected from the two million running words in the whole corpus. The count reflects the distribution of words in the 160,000 word sample.

300. Zipf, G. K. Observations on the possible effect of mental age upon the frequency distribution of words from the viewpoint of dynamic philology. Journal of Psychology, 1937, 4, 239-244.

This is a response to such critics of Zipf's rank-frequency hypothesis as Empson, Joos, and Thorndike. The author points out that the rank-frequency relation holds even when total vocabulary size is different, as in the language of two children of different ages that is discussed in the essay.

301. Zipf, G. K. Homogeneity and heterogeneity in language, in answer to Edward L. Thorndike. The Psychological Record, 1938, 2, 347-367.

Here Zipf replys to Thorndike's criticism of this rank-frequency hypothesis; he then tests his theory by a study of word distribution in several works by James Joyce.

302. Zipf, G. K. The meaning-frequency relationship of words. The Journal of General Psychology, 1945, 33, 251-256.

This study investigates the relationship between the frequency of occurrence of a word and its number of meanings.

303. Zipf. G. K. Human behavior and the principle of least effort. Cambridge, Mass.: Addison-Wesley Press, Inc., 1949.

In this book the author demonstrates his theory of least effort and human behavior in two contexts: language and the structure of personality and human relations; a case of intraspecies balance. Part 1 is of interest to linguists since it deals with the use of language and symbol formation.

304. Zipf. G. K. The psycho-biology of language (an introduction to dynamic philology). Cambridge, Mass.: MIT Press, 1965 (Originally Houghton-Mifflin Company, 1935).

In this book, Zipf explores speech as a natural phenomenon--a biological-physiological and social process--by the use of statistical approaches. He finds that the distribution of words in English approximates an harmonic series. He includes meaning and emotion in his studies of language forms and functions. His book is divided in six chapters: Introduction, Form and Behavior of Words, Form and Behavior of Phonemes, Accent Within the Word, The Sentence (Positional and Inflectional Languages), and The Stream of Speech and its Relationship to the Totality of Behavior.

Aborn, Murray	1, 2
Allen, H. Jr.	144
Allen, J.	3
Allen, W.	4
American Mathematical Society	5
Ardouin, P.	182
Ashen, R.	6
Ashton, E.	7
Auscherman, Marian R.	28
Bailey, D.	8
Bailey, Richard W.	9
Bakaya, R.M.	10
Baker, Sidney J.	11, 12
Bar Hillel, Y.	13
Barber, C.L.	14
Barker, Muhammed	15
Barth, Gilbert	16
Bauelas, Alex	17
Becker, Selwyn D.	17, 18
Beier, E.G.	19
Belevitch, Vitold	20
Belonogov, G.G.	21
Berckel, J.A.	22
Berger, K.	23, 24, 25
Berkowitz, A.	141
Berry, Jack	26, 27
Black, John W.	28
Blankenship, Jane A.	29
Bloch, B.	30
Bond, O.	285
Bongers, H.	31
Booth, Andrew D.	32
Borko, Harold	33
Bourne, Charles P.	34
Bowen, John H.	35
Brader, Marcia	17
Brain, J.L.	36, 37, 38, 39
Buchanan, A.	40
Buchanan, M.	41
Buettner, C.	42
Bull, William E.	43, 44
Burton, Doloren M.	9
Burton, N.G.	45
Bushnell, Paul F.	46
Card, William	47
Carrol, Jean	18
Carroll, John B.	48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58
Carter, Charles W.	101
Chaplin, H.	59
Chomsky, N.	60, 61, 62, 63
Chotlos, J.W.	64
Chretien, D.G.	65
Cole, L.	66
Condon, E.U.	67
Corstius, H. Brandt	22
Daiji, S.	68
Dale, E.	69, 70
Davies, A. (ed.)	71
Davies, P.	57
Denes, P.B.	72
DeVito, Joseph A.	73, 74, 75, 76
Dewey, G.	77

Dingwell, W. 78
Dixson, R. (ed.) 79
Dolby, J.L. 80
Driemann, G.H.J. 81
Durr, William K. 82
Eastman, Carol M. 83, 84
Eaton, Helen S. 85
Eckerson, Lois D. 246
Edmundson, H.P. 86
Elderton, W.P. 87
Eldridge, R.C. 88
Ellegord, Alvar 89, 90, 91, 92
Estoup, J.B. 93
Eyestone, Maynard M. 94
Fairbanks, Helen 95
Flood, W. 96
Ford, Donald F. 34
Fowler, Murray 97
Francis, W. 168
Franklin, H. 98
French Ministry of National
Education 99, 100
French, Norman R. 101
Friedman, E.A. 199
Fries, A.C. 103
Fries, C. 102, 103, 104
Fries, C.C. 172, 173
Frumkina, F.M. 105, 106, 107
Fry, Dennis 108
Fucks, Wilhelm 109, 110
Gammon, Edward R. 111
Garcia Hoz, V. 112
Garvin, Paul (ed.) 113
Gen. Itasaka 136
George, Alexander L. 114
George, H.U. 115
Gerganov, Y.N. 107
Gibson, Tomas 116, 124
Gilmore, T. 117
Good, I.J. 118
Gougenheim, G. 119
Graham, E. 120
Green, J.R. 121
Greenway, P.J. 122
Gross, M. 123
Gruner, Charles R. 116, 124
Guilbert, Louis 124
Guiraud, Pierre 126, 127
Gutkind, A. 289
Harwood, F.W. 128
Hass, W. 278
Haydon, Rebecca E. 129
Hays, D.G. 130
Henmon, V.A.C. 131
Herdan, G. 132, 133,
134, 135
Herold, C.D. 219
Hibbert, H. 136
Hill, Archibald 137
Hill, L. 138
Holstein, A.P. 139
Horn, E. 140
Hornby, A.S. 217
Horowitz, W. 141
Horowitz, M.W. 142
Horton, Iver 250

Howes, D. 143
Hultzen, I. 144
Ichiro, S. 145
Jakobovitz, L.A. 146
Johnson, D.B. 147
Johnson, F. 148
Jones, L.V. 149
Jones, R.M. 150
Joos, Martin 151, 152
Jorden, E. 30, 153
Josselyn, H. 154
Kaeding, F.W. 155
Kaplan, B. 279
Karlgren, Hans 156
Keil, Rolf-Dietrich 157
Kelly, Francis, J. 116
Kibler, Robert J. 124
Kihouka, T. 158
Kochi, D. 159, 160
Koenig, Walter 101
Koutsoudas, Andreas M. 161
Kramsky, J. 162
Kraus, Jirf 169
Krishnomurthy, K.H. 164
Kroeber, Karl 165
Krohn, R. 166
Kublin, H. 167
Kucera, H. 168, 169
Kwasa, S. 117
Lachman, R. 170
Lado, R. 171, 172, 173
Lamb, Sydney M. 174
Lamendella, John T. 58
LeBreton, F. 175
Licklider, J.C.R. 45
Light, Richard L. 176
Loogman, A. 177, 178
Lorge, I. 179, 180, 267
MacMurray, E. 80
MacPhee, E. 40
Machol, Robert E. 161
Mackey, W.F. 181, 182
Malcolm, J. 183
Mandelbrot, Benoit 184
Marchand, H. 185
Marchand, M. 186
Martin, S.E. 59, 187, 188, 189
Maw, J. 190, 191
Mayaji, Hiroshi 192
McCalla, Gorden I 193
McCarus, Ernest 194
McDavid, Virginia 47
McGovern, W. 195
Meddleton, Ivor G. 245
Meir, Helmut 196
Meikle, H. 98
Michea, R. 119
Milic, Louis T. 197
Miller, D.E. 19
Miller, G. 62, 63, 198, 199
Miron, M. 144
Mokken, R.J. 22

Monroe, G. 169
Moore, W. 200
Morgan, B.A. 201
Morris, K. 251
Muller, Charles 202, 203, 204
Nakajima, K. 295
National Institute of Health 205
(The) National Language Research Institue (of Japan) 206, 207, 208, 209, 210
Newman, Edwin B. 199, 211
Newman, J.B. 142
Nice, Margaret Morse 212
Nihonmatsu, R. 54
Nisbet, J.D. 213
Oettinger, Anthony G. 214
Ogawa, Y. 200
Ogden, C. 215
Palmer, H. 216
Palmer, H.E. 217
Perrott, D.V. 218
Petty, W.T. 219
Pfeffer, A. 220, 221, 222, 223
Pimsleur, Paul 224
Plath, Warren 225
Polome, E.C. 226, 227
Posner, Rebecca 228
Pressman, A. 229
Purim, L. 230
Rammuny, Raji 194
Rapoport, Anatol 231
Razik, T. 69
Reed, David W. 232
Reichert, D. 70
Resnikoff, H.L. 80
Richards, Jack C. 233
Richman, B. 57
Rivenc, D. 119
Roberts, A. Hood 234
Robinson, W.P. 235
Rodridguez-Bou, I. 236
Rose-Innes, A. 237
Rosman, E. 240
Rubenstein, Hubert 1, 2
Russo, G.A. 238
Rutherford, R.W. 239
Sacleux, C. 241
Sampson, Jeffrey R. 193
Sauvage, A. 119
Savard, J.G. 181, 182, 242, 243
Scholes, Robert J. 244
Schonell, Fred J. 245
Seashore, Robert H. 246
Sebeok, Thomas A. (ed.) 247
Shannon, C.E. 248
Shapiro, B.J. 249
Shaw, B.A. 245
Sherwood, John 250
Siliakus, H. 251
Simon, Herbert A. 252
Skinner, B.F. 253

Society For International Cultural Relations (Japan) 254, 255
Somers, H.H. 256
Spolsky, Bernard 257
Starkweather, J.A. 18
Stoll, E. 219
Stone, P., et al 258
Strain, J. 98
Swenson, E. 259
Swenson, Rodney 260
Tadashi, Kikuoka 261
Tannenbaum, Percy H. 262
Tarnoczi, Lorant 263
Taylor, G. 264
Thompson, Godfrey H. 265
Thompson, J. 264
Thorndike, E. 179, 266, 267
Traver, A. 104
Uhlirova, L. 268
University of Michigan (English Language Institute) 269, 270
Vakar, Nicholas P. 271, 272
Van, Th. M. 22
Van den Eynde, R. 273
Van Spaandonck, M. 274
Vasilevich, A.P. 275
Voelker, Charles H. 276
Walsh, S. (ed.) 277
Waugh, Nancy C. 211
Wears, M. 239
Wepman, J.M. 149, 278
Werner, H. 279
West, M. 96, 259, 280, 281, 282, 283, 285
Whatmough, J. 286
Whiteley, W. 287, 288, 289
Wijngaarden, A. Van 22
Williams, F. 262
Wilson, P. 290, 291
Winter, Ralph Dana 292
Wisby, R.A. (ed.) 293
Wright, A.M. 128
Wright, C.W. 294
107
Yamagiwa, J. (ed.) 295, 296
Young, I. 297
Zale, E.M. (ed.) 298
Zimmerman, Jon E. 299
Zipf, G.K. 300, 301, 302, 303, 304

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author)
Syracuse University Research Corporation
Merrill Lane, University Heights
Syracuse, New York 13210

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

REPORT TITLE
The Counting of Words: A Review of the History, Techniques, and Theory of Word Counts with Annotated Bibliography.

DESCRIPTIVE NOTES (Type of report and inclusive dates)
Special Report 1 July 1972 - 15 May 1973

AUTHOR(S) (First name, middle initial, last name)
James E. DeRocher, Murray S. Miron, Sam M. Patten, Charles C. Pratt

REPORT DATE
May 1973

7a. TOTAL NO. OF PAGES
295

7b. NO. OF REFS
398

CONTRACT OR GRANT NO.
DAAG-05-72-C-0574

PROJECT NO.

9a. ORIGINATOR'S REPORT NUMBER(S)
SURC TR 73-177

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
NONE

DISTRIBUTION STATEMENT
Approved for Public release, Distribution unlimited.

SUPPLEMENTARY NOTES
NONE

12. SPONSORING MILITARY ACTIVITY
Defense Language Institute.

ABSTRACT

As part of a continuing project of language analysis, SURC presents an essay on the nature and history of frequency counts. The first section deals with the history of such counts and traces them from Early Hellenic times to the present. Section Two is an analysis of techniques used and describes the capabilities and limitations of frequency counts taken in both the English and Foreign Languages. Section Three is an analysis of the statistical lawfulness of vocabulary distributions and presents a comparison and evaluation of the theoretical models used to describe vocabulary distributions. Section Four is an annotated bibliography with an author index provided.

D FORM 1 NOV 65 1473 (PAGE 1)
0101-807-6801

Unclassified
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Structural Analysis Languages Mathematical Linguistics Vocabulary Language Research Descriptive Linguistics Contrastive Linguistics Etymology						

~~900~~